Improving the statistical evaluation of forensic DNA evidence

by

Duncan Taylor

Thesis
Submitted to Flinders University
for the degree of

Doctor of Philosophy by publication

College of Science and Engineering
Discipline of Mathematics and Statistics

31st May 2019

Contents

Declaration:

I certify that this thesis does not incorporate without acknowledgment any material previously submitted for a degree or diploma in any university; and that to the best of my knowledge and belief it does not contain any material previously published or written by another person except where due reference is made in the text.

Personal preface:

Abstract:

The end of the 20th Century saw, in Australia, the beginning of Forensic DNA profiling for use in criminal investigations and Court proceedings. Compared to modern abilities, DNA profiling, when first introduced, had low sensitivity and low powers of discrimination. The type of forensic samples that could be targeted were typically body fluids (such as semen, saliva or blood) that had abundant (at least by today's standards) amounts of DNA available. The laboratory hardware and the profiling systems improved with time and became more sensitive, were able to produce informative results quicker, at less cost and with greater discrimination power. These improvements encouraged the forensic community to branch out from the standard body fluid samples and by the late 1990s forensic samples were being taken from what was termed 'touch DNA', tiny amounts of DNA left behind on a surface, not from a body fluid, but simply from being transferred when the item was touched. These new samples, coupled with the continually increasing sensitivity of DNA profiling, meant that the DNA profiles became more complex, in terms of the number of contributors and the quality and amount of DNA template.

While a substantial level of resources had been expended on the improvement to the ability to generate a DNA profile, a disproportionately small amount of effort had been put into how best to interpret the results. Starting at the turn of the 21st century, a series of methods were developed that could be used to interpret DNA profile. There were two main branches of interpretation methods that formed, which are commonly referred to as the Likelihood Ratio (LR) method (also called the Bayesian method), which dominated in Europe and Australia, and the inclusion probability method (also called the Random Man Not Excluded, or the frequentist method), which dominated in the USA. In the forensic field today the LR method is generally accepted as the superior method and so is the focus of this thesis.

All LR methods have the same foundation, that is, they seek the ratio of the probability of the observed DNA profile (or multiple profiles) given two competing propositions, which typically align with a prosecution stance and a defence stance. The probabilities are assigned for each proposition by taking a weighted sum of all genotype probabilities that apply under that proposition. The simplest form of the Likelihood Ratio method is known as the binary approach, which weights the genotype probabilities with either a 1 or a 0, i.e. they are either included in the sum or they are not. The binary method typically relies on a subjective assessment of the DNA profile by an analyst, who would be utilising a system of rules and threshold for interpretation. There are many shortcomings of such a system, such as; the very restricted pool of profiles to which it could be applied, the inconsistent application between analysts and the waste of much of the information within the DNA profile (i.e. the intensity of each piece of information and its molecular size).

A more elegant approach to weighting the genotypes within the LR approach was termed the 'semi-continuous method'. This method weights the genotypes using probabilities associated with events that occur during the process of generating the DNA profile. The semi-continuous method expanded the types of profiles for which a statistical weighting could be applied and was also able to be applied in a more consistent manner. Semi-continuous systems still did not utilise much information from the DNA profile other than the presence or absence of information and so in that regard still had a limited discrimination powers.

This thesis is a compilation of publications that extend the semi-continuous methods of developing a LR to what has been termed 'fully-continuous'. This is achieved by the use of a much greater level of information from DNA profiles. In order to utilise peaks heights, models have been developed that describe aspects of DNA profile behaviour, that ultimately lead to the patterns of peak intensity seen in a profile. These include models and parameters for stutter, degradation, saturation, peak height variability within and between regions, drop-out and drop-in. For complex DNA profile data, the numbers of combinations that these different parameters can take exceeds the computational ability that would allow an exact solution based on Maximum Likelihood and so a stochastic process using Markov Chain Monte Carlo is developed. The creation of a fully continuous DNA profile interpretation model and a stochastic implementation was trialled of a range of DNA profiles that vary in number of contributors, DNA amounts and degradation levels that might typically be encountered in a Forensic Laboratory.

In addition to the models and systems that allow the deconvolution of complex, mixed DNA profiles, this thesis describes extensions to the LR theory that were developed that allowed a statistical weighting to be provided for the comparison of any reference to virtually any DNA profile. The behaviour of the LR was examined in depth by observing trends in the magnitude of the LR in problems created that varied important factors over a range of plausible values. These trends were aligned with theoretical expectations to judge the performance of the fully continuous system. The system was also extended so that a LR based method could be used to search a database of DNA profiles for either a potential contributor, or a potential relative of a contributor to an unresolvable DNA profile (something that had previously not been possible in the forensic community in Australia).

Methods were developed for calibrating the system to specific laboratories performance so that it provided evidential strengths that were appropriate for the type of data being produced by that specific laboratory. As this concept of expert system calibration, and the concept of a fully continuous system based on a stochastic process, was relatively new in the field of Forensic Biology, some time was spent on validating its performance and instructing others on how they could validate the performance of the system. Validation of the developed fully continuous system was aligned with published guidelines on validation, produced by international advisory bodies on DNA profile interpretation.

A discussion on how the models for deconvolution and LR development could be extended to apply to new situations is provided. Specifically, the deconvolution of DNA profiling data derived from the Y-chromosome (called Y-STR profiling) is shown and the extension of both deconvolution and LR development to consider a range of contributors within the one analysis is given.

To conclude the thesis the work on DNA profile evaluation is placed into a wider case context. This includes a study into the interpretation of the raw electrophoretic data that makes up the DNA profile (and preceding the DNA profile evaluation) and a study into how the support for an individual's presence or absence from a DNA sample can be considered in conjunction with other case and sample information in order to help address queries of questioned activity.

Chapter 1: Introduction

In the years leading up to 2009, the standard method of DNA profile interpretation around Australia involved an analyst (experienced in viewing electropherograms) using their experience-based understanding of DNA profile behaviour to pass judgement on whether an individual of interest could be a contributor of DNA to a sample. Typically, one of three possible opinions would be reached; either the person being compared could be excluded as a contributor, or they could not be excluded as a possible contributor or the complexity was such that no opinion was given. This final opinion was commonly called 'inconclusive', as in 'it is inconclusive as to whether the person could have contributed to the profile', noting that it is in fact always inconclusive as to whether someone contributed to a DNA profile, hence the reason for carrying out an interpretation in the first place. If a person was not excluded as a possible contributor to the DNA profile, then sometimes it was possible to provide a statistical weighting to the opinion. In order for this to occur, the DNA profile had to meet a number of highly conservative rules (referred to as thresholds in the parlance of forensic science) that attempted to mitigate against any possible misinterpretation or overstating of evidential strength. The desire in forensic science is to always bias opinions in favour of falsely stating that a DNA donor is not a contributor of DNA to the sample rather falsely stating that a non-donor is a contributor of DNA to the sample. This desire saw the cultivation of thresholds that were highly wasteful of information in the DNA profile, even when calculating a numerical evidential weight.

The issues with this manual method of interpretation can be grouped into three main categories:

1) The process of applying threshold is binary, i.e. based on the use of a threshold, an event is either deemed possible or impossible. In a DNA profile, an example of this occurs when considering whether two peaks could come from a common donor. Ideally, they should be similar in height but will have been affected by the random variations of peak height, inherent in the process of generating a DNA profile. If a 'balance threshold' were applied it would mean that the possibility that they could be paired was absolute (the probability being assigned as 1) and would remain so for a range of possible peak height values until at some point one peak height (relative to the other) falls below the threshold at which point the pairing becomes impossible (and is assigned a probability of 0). This phenomenon has been described in the forensic community as 'falling off the cliff', where a change in the smallest increment of some measure leads to a diametrically opposing opinion. This is a consequence of applying any threshold and tends to yield a very poor description of reality close to the threshold values. In forensic science, the fact that opposing views could come from a tiny change in situation was exploited by lawyers and the Court as a claimed demonstration of 'untrustworthiness' or 'unreliability'.
2) Due to the logic difficulties associated with thresholds, and the desire for conservativeness the threshold used meant that much data within the DNA profile was wasted. Apart from it being typical to use only the presence or absence of peaks in a LR calculation (and not utilising their heights other than the initial, manual,

pre-calculation process of exclusion or non-exclusion) it was also typical to ignore entire blocks of information (termed 'loci' in a DNA profile and referring to one of the regions of the DNA targeted by a DNA profiling system) or use approximations under the belief that they were always conservative.

3) Because the initial assessment of exclusion or non-exclusion was largely experience based analysts in different laboratories, or different analysts within the same laboratory, or even the same analyst at different times, would come to different conclusions on the same profile. Even when initially trained in the same manner, if one analyst happened to come across an aggressive defence in court that challenged their opinion of non-exclusion, then they would be more likely to shy away from a non-exclusion opinion in the future. The forensic community tried to rectify the situation via a system of thresholds, but found they had to have a balance: Too simple and the system would not be applicable to many profiles and following the rules would find an analyst in areas of undefendable logic-traps. Too complex and they could not be applied in a consistent manner, rendering pointless the very reason they were created.

In 2009 one forensic laboratory in Australia came to the realisation that the manner in which they had been applying their system of thresholds was not acting conservatively in the way they intended. In very short time they altered the manner in which they were carrying out their evidence evaluations and reissued reports to the courts in a number of cases. Many other factors came into play, however the outcome was that the courts lost faith in the results they were given and the biology section of the forensic laboratory was temporarily shut down pending a review. In late 2009 a 'crisis meeting' was held with attending representatives from each forensic biology laboratory around Australia and New Zealand to address the issue. Coming out of the crisis meeting were two main points:

1) Laboratories realised how differently they were evaluating DNA profile evidence. This was most clearly demonstrated when an exercise was conducted whereby a series of DNA profiles were sent around to each laboratory for assessment, and the results later compared and contrasted.
2) Laboratories realised the need for an Australia-New Zealand statistical specialist working group (Stats SWG) with the overarching remit of standardisation and education. The formation of this group in 2010 saw John Buckleton as chair and Duncan Taylor as vice chair (who then become chair of the group in 2012, until 2014)

The formation of the Stats SWG was one of the biggest positive moves made towards DNA evidence evaluation for many years in Australia, as it allowed free discussion of ideas and concepts between laboratories. While, in the short term a more standardised threshold-based system was settled on (elements of which are still refer to today), in the longer term the group agreed that the best outcome would be to move from threshold-based systems altogether. Work was started on a system of DNA evidence interpretation that replaced rules with models, and threshold with distributions. Under the guidance of two forensic organisations (Forensic Science SA and the Institute of Environmental Science and Research in New Zealand) and the

National Institute of Forensic Science in Australia (headed at that time by Alastair Ross), work was started on developing a system that could interpret mixed DNA profiles originating from two individuals, taking only the most fundamental of profile aspects into account, with the potential that it could later extend to more complex problems. Over the next year formulae were derived, and systems developed that could address the issue and eventually a system based on Markov Chain Monte Carlo was developed. The mathematics was derived in generality, so that there was no limit to the profile complexity that could be analysed (save perhaps computing power, time and peoples comfort levels). By 2011, a basic working system had been developed that passed an initial trial, by testing a series of mixed DNA profiles produced for validation. By mid-2012 STRmix™ was introduced into active casework in South Australia and New Zealand, with the rest of Australia coming on board over the following few years.

Chapter 1 provides a brief introduction to the basics of DNA profiling (section 1.1) and interpretation (section 1.2), which is included so that this thesis can be read as a closed piece of work, even for those who are unfamiliar with the general topic. Section 1.3 provides some content on the continuous system of DNA profile interpretation that is the core of the thesis.

Chapter 2 elaborates on the models that describe DNA profile behaviour. Once defined (and tested) these models can then be combined into a complete system used to describe, probabilistically, observed DNA profiles, and more importantly assess the potential contribution of nominated individuals to evidence profiles.

Chapter 3 describes the statistic used to evaluate DNA evidence, the likelihood ratio (*LR*). Within the LR there is a vast array of topics that are considered; the way propositions are formulated, the parameters within the LR model that contribute to uncertainty, the consideration of complexities such as the presence of related individuals in a DNA profile, and the exchange between considering a single component of a mixture to the entire mixture. Many of these topics were highlighted during the development and testing of STRmix™ and required the development and derivation of mathematical descriptions and solutions.

As STRmix™ uses models that describe DNA profile behaviour, it became apparent that the system would need to be calibrated to each laboratory process to which it was applied. Chapter 4 discusses the calibration of a complex MCMC system to the functioning of a laboratory.

Chapter 5 presents a series of works that were required to test the functioning of the MCMC system against theoretical expectations, in an effort to demonstrate its ability to provide appropriate evidential strength when evaluating individuals' potential contribution to a mixed DNA profile.

Chapter 6 demonstrates the extent to which STRmix™ has been successful in achieving consistency between analysts and laboratories.

Chapter 7 and 8 present work focussed on the future directions of the system of DNA evidence evaluation (chapter 7) and then efforts to improve the data being produced prior to the use of STRmix™ and after STRmix™ (chapter 8).

Chapter 9 provides information on the impact that this work has had on the forensic community, mainly from its use in the DNA analysis software STRmix™

1.1 An introduction to DNA profiling

One of the most commonly used forensic science disciplines is forensic biology. Typically, when a criminal case is submitted to a forensic biology institution, it is done so with the aim of identifying whose DNA is present on an item, which can be used as evidence that someone has (or has not) been in contact with an item of probative interest in the case. In order to carry out the task of identifying DNA on an exhibit, the exhibit is examined, and samples are taken from areas of interest (e.g. the handle of a weapon to identify who may have handled it, the collar of a t-shirt to identify who may have worn it, or an intimate swab taken from a victim to identify who may have been in contact with them). The sample then has cellular material broken open to release DNA (a process known as DNA extraction), the DNA is quantified and then targeted segments of the DNA are copied millions of times, with each fragment having a fluorescent tag attached, in a process known as polymerase chain reaction (PCR). The amplified DNA fragments are separated according to size using a capillary of acrylamide gel, and then detected by excitation of the fluorescent tags and detection by a charged couple device camera. The greater the number of DNA strands that were present in the initial sample, the more amplified PCR fragments will be produced and the greater the detected fluorescent signal. The resulting graph of fluorescence over time is referred to as a DNA profile.

Figure 1 shows a DNA profile from a single individual, created using a commercially available PCR kit called GlobalFiler™ (Thermofisher). GlobalFiler™ targets 24 regions of human DNA, two of which are associated with determining the sex of an individual and 22 of which are highly mutable short tandem repeats (STRs) that are used for individualisation.

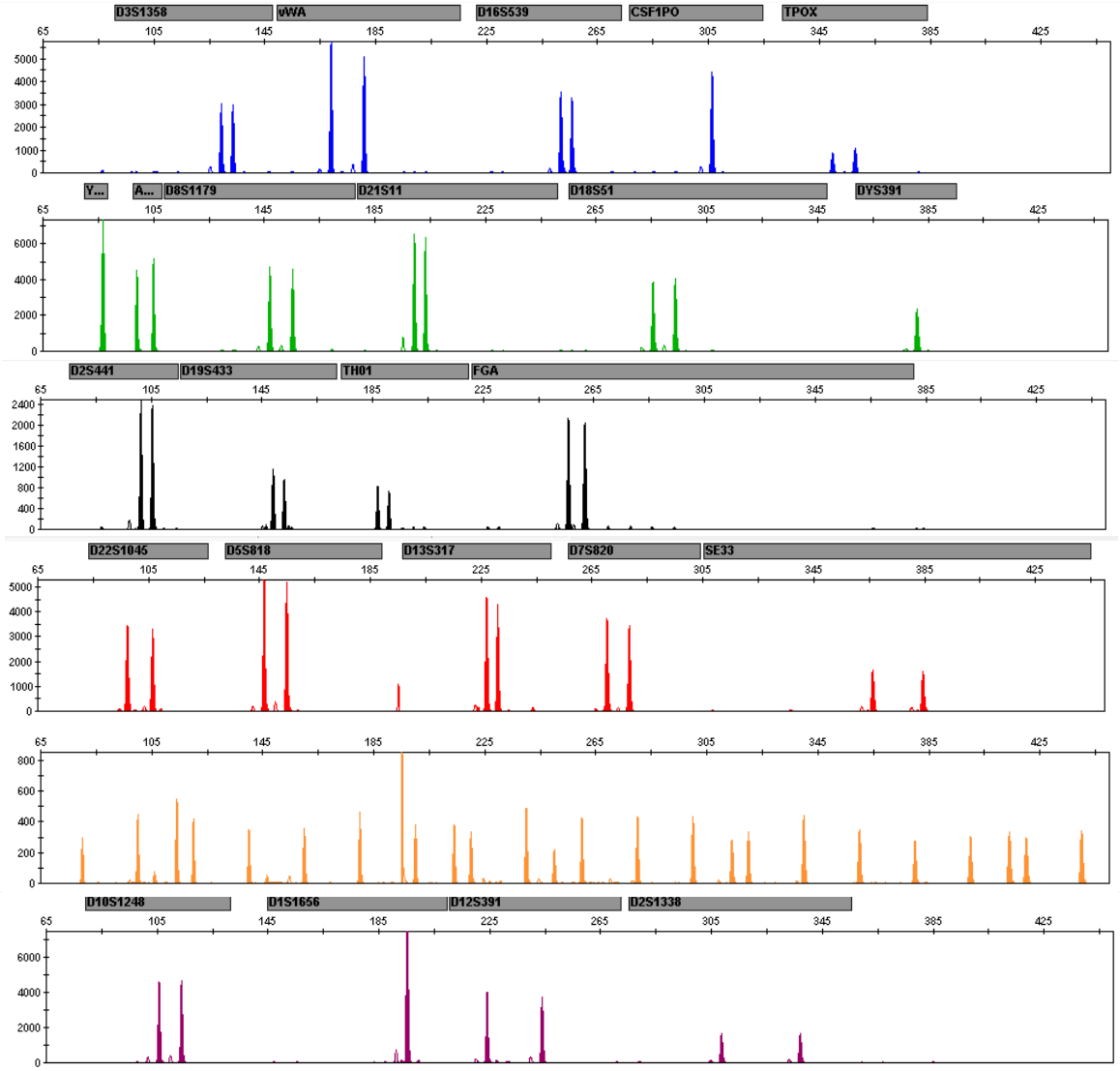


Fig. 1. A DNA profile shown in the form of an electropherogram (epg)

The horizontal axis represents the molecular weight of the PCR amplified fragment, expressed in base pairs. The signal is divided into six horizontal panels, which represent the six types of fluorescent tag used during the PCR process (6-FAM, VIC, NED, TAZ, LIZ and SID). The vertical axis is measured in relative fluorescent units (rfu) and represents the amount of starting DNA in the sample.

Once the fluorescent signal has been captured from the capillary electrophoresis instrument it must be interpreted before the information can be evaluated in respect to a criminal matter. The interpretation consists of designating areas of fluorescence on the DNA profile into categories. Some of these categories are useful in the evaluation as they represent information about the DNA that was present on the originally sampled exhibit. Other types of fluorescence are artefactual and arise as a consequence of producing the DNA profile. The main types of fluorescent signal that are classified are:

- Baseline – The level of background noise present in DNA profiles produces by the capillary electrophoresis instrument. Baseline is not used in DNA evidence evaluations.
- Allele – The peaks that represent the STR fragments that have been amplified during the PCR process. Alleles are used in all evidence evaluations.
- Stutter – Artefacts produced during the PCR process, stutters are replication errors that lead to small peaks that appear around the 'parent' allele. Stutters can be of different types commonly named in relation to the position relative to the parent peak i.e. 'back stutter' (one STR unit less than the parent), 'forward stutter' (one STR unit greater than the parent) or 'half-stutter' (half a STR unit less than the parent). Stutters are used in some evidence evaluations (depending on the sophistication of the model being used for evaluation).
- Pull-up – Artefacts due to the overlap of the distribution of wavelength emitted by each of the fluorophores used in commercial profiling kits. When many fragments labelled with a specific fluorophore are detected by the CCD camera a high intensity peak is produced in the corresponding dye lane. Lower intensity peaks are seen in dye lanes that correspond to fluorophores with similar excitation wavelengths. Pull-up is not used in DNA evidence evaluations

Once a DNA profile has been generated in the laboratory and processed to remove any unwanted artefactual signal, it can be used for evaluation. Evaluation is the name given to the comparison of the evidence profile to reference DNA profiles, to determine whether the donor of the reference could also be a donor of DNA to the evidence. The weight of evidence that is calculated to carry out this task is a likelihood ratio, which is explained in many places throughout this thesis, and which is introduced in more detail in section 1.2.

1.2 The evaluation of DNA profile data using the likelihood ratio

The standard method of modern DNA profile evaluation is the use of a likelihood ratio (LR). The LR considers the probability of obtaining the observed data (O) given two competing propositions which align with the prosecution (called Hp) and defence (called Hd). The LR is then:

$$LR = \frac{\Pr(O \mid Hp)}{\Pr(O \mid Hd)}$$

In order to evaluate this expression, a series of nuisance parameters must be considered. The parameter that has the longest recognition are the genotypes of the contributors to the DNA profile (call the set of genotypes that could describe the profile S, and there are J of them to consider). The LR is then stratified across genotype sets:

$$LR = \frac{\sum_{j=1}^{J} \Pr(O \mid S_j) \Pr(S_j \mid Hp)}{\sum_{j=1}^{J} \Pr(O \mid S_j) \Pr(S_j \mid Hd)}$$

where the term $\Pr(O \mid S_j)$ is the probability of obtaining the observed data if genotype set 'j' describes the genotypes of the underlying contributors (and are often referred to as weights) and $\Pr(S_j \mid Hx)$ is the probability of that genotype set, given the proposition (and the genotypes of the contributors specified within it). The great challenge to the forensic community in DNA profile evaluation is the ability to assign values to the weights. This point (and more detail on LRs) is discussed in detail in various chapters of the thesis.

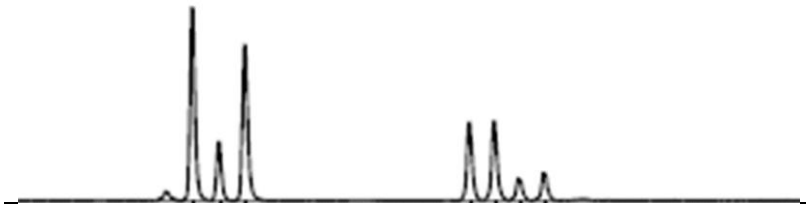
Prior to the advent of modern software systems that can be used to analyse DNA profile data (these will be discussed in depth during the thesis) the method of DNA profile interpretation and evaluation was a threshold-based system. The use of threshold was mentioned the introduction as assigning the values of the weights in the LR as 0 or 1. It is worth briefly describing the threshold-based system of DNA profile interpretation because:

- It sets the scene that led to the development of probability-based DNA profile interpretation systems (discussed throughout this thesis)
- The probability-based DNA profile interpretation systems utilise models that remove the need for these thresholds
- Once the thresholds have been applied to the DNA profile, the resulting 'filtered' information is then used in the likelihood ratio evaluation

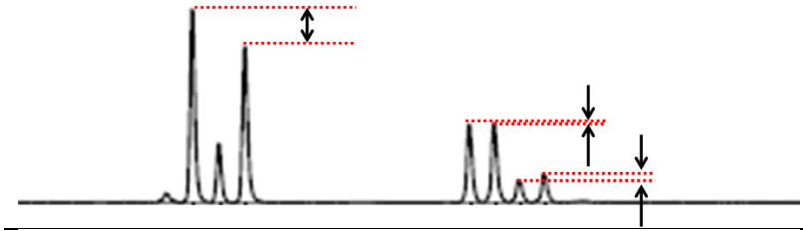
Figure 1.2.1a shows two regions of a DNA profile that an analyst may wish to interpret. Say that the analyst has decided that the profile originates from two individuals and wants to 'interpret' the various genotypes that could give rise to this combination of peaks. This would typically occur by the application of a series of thresholds to screen out potential combinations of the allele as impossible (more accurately, improbable to the point that they were going to be discounted as possible genotype combinations). Panel b in Figure 1.2.1 shows the application of a peak balance threshold. In this example, consider that the analyst

wishes to interpret the genotype of the major contributor. To this end the two tallest peaks at each locus are considered as potentially pairing, and belonging to the major DNA donor. If their height is within a predefined threshold of acceptable balance this would be an allowed pairing and the analyst could then continue to apply additional thresholds. If they fall outside the balance threshold, then the analyst would abandon the genotype set being considered. Panel 1.2.1c shows the application of a dropout threshold (in blue) and a stutter threshold (in green). In later chapters in this thesis these terms will be explained and modelled in great detail and so that work is not repeated here. It will suffice to say that both dropout and stutter are properties of DNA profiles, for which some behaviour is expected and for which thresholds have (in the past) been set. Again the analyst would consider the genotype and the combinations of dropout or stutter that the genotype being considered would require in the observed data in order to make a decision as to whether this genotype could reasonably describe the observed data, and again this may eliminate the genotype set from consideration. The final threshold shown in Figure 1.2.1 id panel d which applies a mixture proportion threshold (in essence the purpose of this threshold is to ensure that the larger peaks at one region would align with the larger peak at another region, as expected by DNA amounts being contributor specific and constant across regions). Again this may eliminate or allow genotype sets through the interpretation.

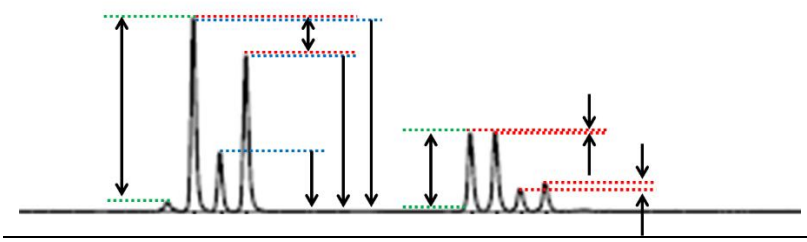
At the conclusion of the process the analyst would left with genotype sets that passed all the threshold-based rules, and if there were only one of these then it would be possible to interpret that component of the profile distinctly from any other. This process of threshold-based interpretation was very wasteful of information, difficult to apply consistently and would often lead to situations where no statistical calculation could be conducted (as too many genotype sets were considered as possible).



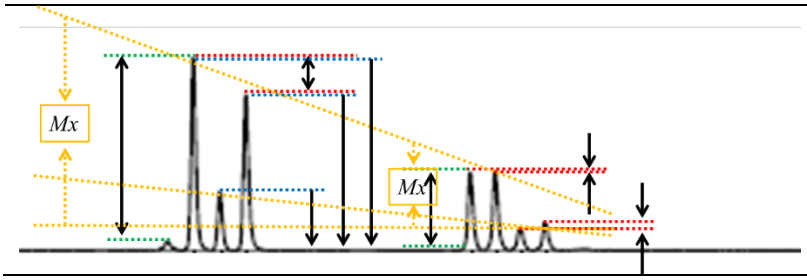
a) DNA profile



b) Application of peak balance threshold (red)



c) Application of dropout threshold (blue) and stutter threshold (green)



d) Application of mixture proportion threshold (yellow)

Figure 1.2.1: Example of a threshold-based system of DNA profile interpretation

As understanding of DNA profile behaviour and interpretation methods grew over the years, it became possible to apply models to profile data, rather than thresholds. For example, instead of applying a balance threshold, a shift was made to a sliding scale of probability i.e. this level of imbalance is only seen in X% of paired peaks. These models then were refined and eventually used in probabilistic genotype interpretation systems. It is the transition from the manual, threshold-based systems of DNA profile interpretation to the computerised application of statistical models to assign values to the weights in the LR that is the core of this thesis and whose aspects are described in detail in the chapters that follow.

1.3: The fully continuous Bayesian interpretation method

Book chapter: 'Chapter 9: The Continuous Model', written by Duncan Taylor, Jo-Anne Bright and John Buckleton. From the book 'Forensic DNA Evidence Interpretation' Second edition. Edited by John Buckleton, Jo-Anne Bright, Duncan Taylor. CRC Press. 2016. – *Book cited 291 times*

Statement of novelty: This chapter is a new chapter, not present in the original edition of the book. The majority of material in this chapter is new, or summarises work carried out either solely or in conjunction with colleagues. It has been written in a manner that is generally more accessible to individuals who are new to the field of forensic statistics than the original description in the scientific papers from which they derive.

My contribution: My contributor was as main author. I initiated the writing of this chapter and contributed a majority of the work within.

Research Design / Data Collection / Writing and Editing = NA / NA / 60%

Additional comments: This chapter is a gentle introduction to the idea of fully continuous interpretation systems and briefly summaries many points in other works I present within my thesis.

9

The Continuous Model

Duncan Taylor, Jo-Anne Bright, and John S. Buckleton

Contents

Introduction

The key point of difference of a continuous model[741–744] is that it considers the peak heights as a continuous variable. This chapter describes the biological models and statistics behind STRmix™, one such continuous model.

Let the genotype of person i be G_i. The genotype of the person of interest is G_P. To form the likelihood ratio (LR) we consider two propositions H_p and H_d, chosen to align with the prosecution and the defence, respectively. We consider a mixture assigned as coming from two people although the principle is quite general. The two propositions H_i therefore each define one or many sets of two genotypes G_i and G_j as the proposed contributors. These may or may not include the person of interest (POI). Typically the POI is included in all sets under H_p, but not under H_d. There may or may not be genotypes from other persons known or reasonably assumed to be part of the mixture.

We seek

$$LR = \frac{p(G_C \mid H_p, G_S, I)}{p(G_C \mid H_d, G_S, I)}$$

For simplicity we drop the background information I from the conditioning henceforth and use p for a probability density and Pr for a probability. It is convenient from here on to consider terms of the type $p(G_C|S_k,G_P)$. Let H_i specify the sets of pairs of genotypes S_k $k = 1 \ldots M$, then

$$p(G_C \mid H_i, G_S) = \sum_{k=1}^{M} p(G_C \mid S_k, G_S) \Pr(S_k \mid H_i, G_S)$$

The binary model assigns the terms $p(G_C|S_k,G_P)$ the value 0 or 1 depending on whether the crime profile is deemed possible or impossible if it originated from the genotypes specified by S_k.

If one sample is run multiple times the results will not always be the same. Both the absolute and relative peak heights may vary within and between loci.[118,745] What the binary model is doing is assigning the values 0 and 1 based on very reasonable methods that approximate the relative values of $p(G_C|S_k,G_P)$. In essence $p(G_C|S_k,G_P)$ is assigned as 0 if it is thought that this probability density is very small relative to the other probability densities. It is assigned a value of 1 if it is thought that this value is relatively large. As such it is an approximation.

This method has served well for a number of years and in a great many cases. However all approximations suffer from some loss of information.

A fully continuous model for DNA interpretation is one which assigns a value to the probability density $p(G_C|S_k,G_P)$ based on treating the peak heights as a continuous variable.

Such models may require some preprocessing, say of stutter peaks, or may be fully automated. These methods have the potential to handle any type of non-concordance and may assess any number of replicates without heuristic preprocessing and the consequent loss of information. Continuous methods are likely to require models for the variability in peak heights and potentially stutter.

Many of the qualitative or subjective decisions that the scientist has traditionally handled – such as the designation of peaks as alleles, the allocation of stutters and possible allelic combinations – may be removed. Instead the model takes the quantitative information from the electropherogram (epg), such as peak heights, and uses this information to calculate the probability of the peak heights given all possible genotype combinations.

Profile Information

As described in this book the DNA profile evidence is typically assessed in the framework of an *LR*. *LR*s have the general form

$$LR = \frac{\sum_i w_i \Pr(S_i \mid H_p)}{\sum_j w_j \Pr(S_j \mid H_d)}$$

where w_x is a weight for a set of explanatory genotypes (S_x). In Chapter 8 we showed how the weights can be a list of ones or zeros in a binary system or dropout and drop-in probabilities in a semi-continuous system. Both these systems summarize the DNA profile data in some way. In the case of a binary system the summary comes in the form of interpretational thresholds. Peaks are summarized by grouping them into binary categories of either passing or failing a threshold-based test. For example, a dropout threshold (ST) will group single-peaked loci either as originating from a homozygote or potentially having a dropped-out partner peak. Semi-continuous systems summarize the data in that they use the DNA profile to develop probabilities for dropout or drop-in that are applied to all peaks and all posited contributing genotypes. Semi-continuous systems may also have a secondary system for screening out potential contributing genotypes based on stutter or heterozygote balance (Hb) thresholds.

The more correct and relevant information a system is able to make use of, the better its ability will be to differentiate true from false donors.[746] Figure 9.1 shows diagrammatically that as more information is provided to a DNA profile analysis system (either with more DNA, more polymerase chain reaction [PCR] replicates, simpler profiles or more information about assumed contributors) the ability to distinguish true from false hypotheses is increased.

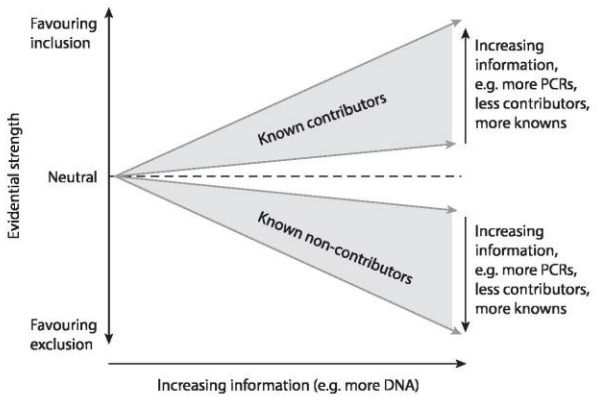


Figure 9.1 Demonstration of the effect that information has on the ability to distinguish true from false hypotheses.

This same thinking can apply to the system being used to interpret a DNA profile, i.e. the more relevant information that is provided, the better the system will be able to distinguish true from false hypotheses. In simple terms true contributors will have more support for their inclusion to a profile and non-contributors will be excluded more often or with more strength. This situation then leads to the question, what information exists within a DNA profile?

On face value the observable data are as follows:

- Peak heights
- Molecular weights
- Allelic designations (which broadly represent underlying DNA sequences)

However there is further information that can be provided to an interpretation system that is not directly obtained from the DNA profile:

- How peak heights are related to template DNA amount
- How DNA profiles degrade
- How loci amplify at differing efficiencies
- How peak heights have a level of variability
- Information about the generation of artefacts during PCR (i.e. stuttering or drop-in)
- How replicate amplifications of the sample extract behave

All of these are termed *models*. There is a further class of information that could be described as *calibration data*. This will typically be specific for a set of laboratory hardware and a method of profile generation:

- How much a specific allele and locus combination is expected to stutter
- How much peak heights are expected to vary (both allelic and stutter)
- How much peak height balance is expected between loci
- How much drop-in is expected
- When a capillary electrophoresis instrument is expected to reach fluorescent saturation
- Below what level a laboratory is not prepared to read information (typically referred to as a baseline or analytical threshold)

It is apparent that both the list of models and the calibration data list is typically the type of information learnt by analysts that interpret DNA profiles. Their knowledge of these DNA profile behaviours and laboratory performances has classically been used in their assessment of DNA profiles prior to interpretation in an analytical system.

The final set of data that could be provided to a DNA interpretation system would relate to specifics of a DNA profile being analyzed. We term these the *unknowables* as in reality their true value can never be known:

- The number of contributors to a profile (N)
- DNA amounts of each contributor (t_n)
- The degradation of each contributor (d_n)
- The amplification efficiency of each locus (A^l)
- Replicate amplification strength (R_r)
- The level of peak height variability within the sample

There are two points to note here. First there is typically a model that describes the behaviour of each of the unknowables. Second again in classic DNA profile interpretation an analyst would be trialling different combination of these factors (perhaps unknowingly) in their assessment of potential contributing genotypes.

Continuous DNA profile analysis systems seek to make use of all the observable data using models and calibration data and even the unknowable information in order to deconvolute a profile into a list of genotype sets (S, a set of N individual genotypes) each with an associated weight that in general terms describes how well the observed data is described by that particular genotype set.

STRmix

There are at least four implementations of the continuous approach of which we are aware.[110,111,743,747] For a good summary of the programmes and their abilities we direct the reader to Ref. 462. While it would clearly be more balanced to discuss the various continuous approaches available at this time, we are unable to do justice to three of them due to a lack of in-depth understanding and in some cases suggestions of potential litigation. We therefore concentrate on the one with which we are most familiar.

STRmix uses standard mathematics and a model for peak and stutter height. Total allelic product modelling assumes a degradation curve that is exponential but allows each contributor to a mixture to have different curves. Individual loci are allowed a limited liberty to be above or below this curve. The total allelic product is split into stutter and allelic peaks using allele-specific stutter ratios developed from empirical data. The contribution to a peak from different allelic or stutter sources is assumed to be additive. The variability about the expected peak height is modelled on empirical data and the relative variance is large for small peaks and small for large peaks. This variance is allowed some limited flexibility within the system so that it can adapt slightly for good or bad profiles. Independence is assumed across peaks at a locus and between loci.

The model for the peak and stutter heights and other assumptions results in a probability density for the profile given a set of input parameters for such things as template and degradation. These input parameters are unknown.

To deal with these unknown inputs STRmix uses Markov chain Monte Carlo (MCMC) and the Metropolis–Hastings algorithm. These terms will be unfamiliar to most forensic biologists but appear to be teachable. MCMC is close to a 'hot and cold' game.

It is likely that continuous methods will supersede all other methods. At this stage we suspect that STRmix has a slightly higher false exclusion rate than the semi-continuous models and a much lower false inclusion rate. The false exclusions for STRmix are usually caused by unusual PCR behaviour.

The software has improved the number of interpretable volume crime cases in New Zealand by 17%.

Mass Parameters

We may consider the evidence of the crime stain G_c to consist of a vector of observed peak heights $\mathbf{O}$ made up of a number of individual observed peak heights O^l_{ar} for allele a at locus l for replicate r. Let there be R replicates and L loci.

To describe these observed peaks we must consider various values for the unknown parameters. We introduce parameters to describe the true template level. Experience and empirical studies suggest that the height of peaks from a single contributor are approximately constant across the profile but generally have a downtrend with increasing molecular weight. Given this general downtrend individual loci may still be above or below the trend. In addition the slope of the downtrend trend may vary from one contributor to another. The product from the amplification of an allele is dominated by correct copies at the allelic position and back stutter at one repeat shorter than the allele. There are a number of other more minor products ignored

in this treatment. We term the sum of the allelic and back stutter product *total allelic product*. We require a term for the true but unknown template level available at a locus for amplification. This is a function of the input DNA and any degradation or inhibition effects. Since template is described by weight, usually in picograms, we coin the term *mass* to subsume the concepts of template, degradation, inhibition and any other effect that determines the expected total allelic product at a locus.

Hence the mass of an allele at a locus is modelled as a function of various parameters which we collectively term the *mass parameters*. These are as follows:

1. A constant t_n, for each of the n contributors that may usefully be thought of as template amount.
2. A constant d_n, which models the decay with respect to molecular weight (m) of template for each of the contributors to genotype set S_j. This may usefully be thought of as a measure of degradation.
3. A locus offset at each locus, A^l, to allow for the observed amplification levels of each locus.
4. A replicate multiplier R_r. This effectively scales all peaks up or down between replicates.

We write the mass variables $\{d_n: n = 1, \ldots, N\}$ and $\{t_n: n = 1, \ldots, N\}$ as **D** and **T**, respectively, $\{A^l: l = 1, \ldots, L\}$ as **A** and $\{R_n: n = 1, \ldots, R\}$ as **R**. The variables **D**, **A**, **R** and **T** are written collectively as **M**.

Template

The heights (or areas) of the peaks within the epg are approximately proportional to the amount of undegraded template DNA.[748–751] Therefore when building a picture of an expected profile the amount of DNA for each contributor will directly relate to the peak heights of contributors. We show this empirically by calculating the average peak height for GlobalFiler™ (Thermo Fisher Scientific, Waltham, MA) profiles generated using varying amounts of DNA (Figure 9.2).

As expected there exists some stochastic variation in average peak heights; however a clear linear relationship can be seen between input DNA and fluorescence.

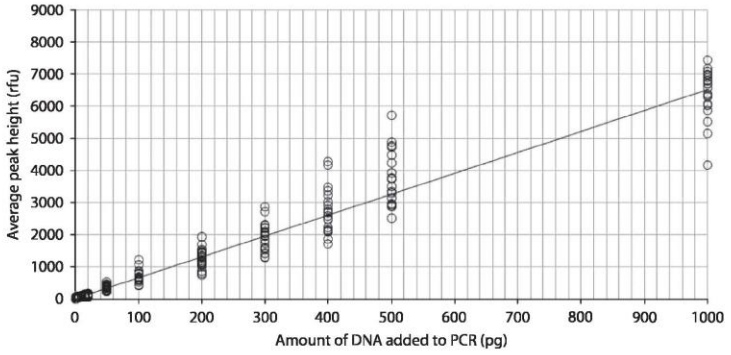


Figure 9.2 Average peak height for GlobalFiler profiles produced on a 3130*xl* with varying amount of input DNA.

It is also expected that peak heights are additive; i.e. if there are multiple sources of a single allele, the height of that allele will equal the sum of the individual expected heights from each source. This is termed *stacking* in the United States. This additivity is assumed to hold true whether the sources are all allelic or whether they are a combination of stutter and allele.

Degradation

The heights (or areas) of the peaks within the epg are approximately proportional to the amount of undegraded template DNA.[748–751] However this relationship is affected by a number of systematic factors. Notable among these factors is the molecular weight (m_a) of an allele, a.

A typical epg has a downward trend with increasing molecular weight. This is variously described as the degradation slope or the 'ski slope'.[729,752,753] The term *degradation slope* alludes to a suggested cause, degradation of the DNA. There are many chemical, physical and biological insults which are believed to contribute to DNA degradation or inhibition of a profile. Environmental factors such as humidity,[754] bacteria[753] or other forces such as ultraviolet light break down the DNA, destroying some fraction of the initial template.[755] Although the cause of the slope may not be known, we will refer to this ski slope effect as *degradation* to comport with common usage. Of interest is that fresh buccal scrapes processed immediately show a degradation slope.

It is important to understand how degradation affects these models. The simplest model is linear. That is, the expected peak height declines constantly with respect to molecular weight. This can be demonstrated crudely by taking a paper epg and drawing a downward sloping straight line across the apex of the heterozygote peaks from the lowest molecular weight locus to the highest molecular weight locus.[120]

If the breakdown of the DNA strand was random with respect to location, then we would expect that the observed height of peak a, O_a, would be exponentially related to molecular weight.[728]

Amplification Efficiency

When template DNA amount, degradation, fluorescence and peak height variation are taken into account using the models described thus far, the variation in peak heights between loci is still more variable than predicted. This additional variability arises from differences in amplification efficiencies between loci. These differences appear to vary with time and maybe even sample to sample. They may be affected by, at least, co-extracted materials that affect the PCR process. Imagine, initially, that we allow a locus-specific amplification efficiency, A^l (LSAE), at each locus. Consider the profile in Figure 9.3.

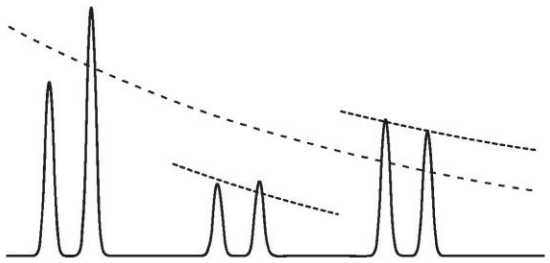


Figure 9.3 Profile showing expected heights at a whole profile level with the dashed line and the adjustment made by locus-specific amplification efficiency if it were completely free to take any value.

If we allow the A^l variables to be free, we will model the mass of template at each locus according to the dotted lines in Figure 9.3. Look now at the heterozygotic loci. If one allele is above the mass line then the other is always below. In fact if one is x rfu above then the other is x rfu below. They are fully correlated. This is because the mass has been overfitted to the data. How could we fit at the correct level, neither over nor underfitted?

Consider the question, how high is peak b? It would be helpful to know the height of peak a, the partner allele of b from a heterozygote. Let us say peak a is height 400 rfu. It would make sense to guess that peak b was also about 400 rfu. However further imagine that you were told that peak a was a bit high relative to the true template at this locus. In fact the true template at this locus suggests that peak a should be of height 350 rfu and in this case it must have varied upwards. Given this knowledge, we would guess that peak b would be about 350 rfu. What this suggests is that if we know the true template at a locus then the height of peak a is not further information regarding the height of peak b. Another useful way to think of this is that the two peaks of a heterozygote should scatter around the true template and if one goes up that does not imply the other one will. This gives us a diagnostic to find the true template. It is that value where the two peaks of a heterozygote are uncorrelated when we consider their variation from this value.

To obtain this value we cannot allow the A^l values to be free. If we did they would overfit.

The continuous method with which we are most familiar is STRmix. We would write in more generality if we knew better what the other implementations do. However it is very likely that the principles are the same. The method applied in STRmix to find the true template at a locus allows the A^l values some limited freedom to fit to the data. The amplification efficiency at each locus is modelled by a lognormal distribution with a mean of zero and a fixed, but optimized, variance determined from laboratory calibration data:

$$\log_{10}\left(\frac{O\{A^l\}}{E\{A^l\}}\right) \sim N\left(0,\sigma_A^2\right)$$

where $O\{A^l\}$ is the observed amplification efficiency for locus l. Note that $E\{A^l\}$ is the expected amplification efficiencies, which we expect to be 1, giving $\log_{10}\left(A^l\right) \sim N\left(0,\sigma_A^2\right)$ where we have just used A^l to signify the observed amplification efficiency for locus l.

By applying this penalty (strictly a prior distribution) there is a pull back towards the mean. This acts against peak height variability like rubber bands pulling the expected heights for the locus back towards the expected level for the whole profile (Figure 9.4 shows this process diagrammatically).

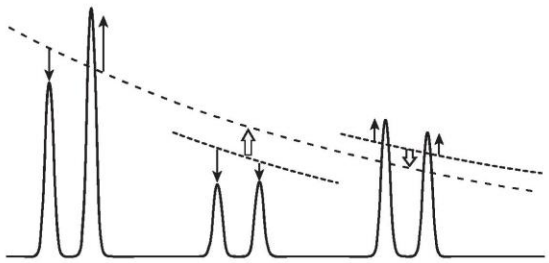


Figure 9.4 Profile showing expected heights at a whole profile level with a dashed line, the influence of peak height variability (solid arrows) pushing the locus-specific expected height (dotted lines) towards the observed heights and the effect of locus-specific amplification efficiency (A^l) (hollow arrows) pulling the locus-specific expected heights back towards the whole profile-expected heights.

Replicate Amplification Strength

Replicate amplification efficiencies are used in the calculation of total allelic product and scale all peaks up or down relative to one another. This allows the inclusion of multiple PCR replicates into a single analysis even if different amounts of template DNA have been added to the PCRs, as long as they originate from the same DNA extract.

The simplest model for replicate amplification efficiencies makes the assumption that the DNA profiles will contain the same individuals with the same degradation, the same relative DNA amounts and the same locus amplification efficiencies.

Note that the replicate amplification terms should be scaled by their logs. In other words, if two replicates were used in an analysis and one was four times the intensity of the other, then the appropriate replicate amplification efficiencies (to prevent differences when considering the two PCRs in different orders) would be 50% and 200% rather than 100% and 400%.

Generating Weights

In order to generate the weights we start with the observed evidence (**O**), which will be a number of peaks at a number of alleles (a), across a number of loci (l) and replicate amplifications (r). Individually each peak can be referred to by $O^l_{a,r}$. We seek the following:

$$\Pr(\mathbf{O}\,|\,I)=\Pr\left(O^1_{1,1}\ldots O^L_{A,R}\,|\,I\right)$$

where I is all the background information the system has about DNA profile behaviour and laboratory calibration. From this point on we omit the I term but recognize that it is important to the calculation. To progress this further we require information about the mass parameters. These of course are unknowable for a DNA profile and so it appears initially as though an impasse has been reached. However in order to consider the effect of mass parameters on the probability of the evidence we need not know their value, rather just the effect that each of the mass parameters has on the probability of the observed data. This allows them to be considered 'nuisance' parameters and integrated over without ever knowing their true value. Box 9.1 explains the concept of integrating out a nuisance parameter.

BOX 9.1 INTEGRATING OUT A NUISANCE PARAMETER

Imagine that we were interested in the average foot size within a population of people. However this information was not readily obtained by records and measuring foot size was impractical. Records were however present for the distribution of height in the population and additionally studies had been carried out on the link between foot size and height.

We define some terms:

F = foot size

H = height

We create a model linking foot size to height:

$$F = E[F|H = h]$$

We seek the average foot size. We could initially consider a simplified model where the average height was calculated ($\bar{h}$) and then calculate the foot size by the following:

$$\hat{F} = E\left[F\,|\,H=\bar{h}\right]$$

Doing this makes a number of simplifying assumptions about the distribution of heights and the model linking foot size and height. If we wanted to take the distribution heights into account (and use a model that better described reality) we could start splitting

height into brackets, for example if it were split into two brackets around 150 cm, and take the average of each bracket ($\bar{h}_{>150\,\text{cm}}$ and $\bar{h}_{<150\,\text{cm}}$):

$$\hat{F} = E\left[F \mid H = \bar{h}_{>150\,\text{cm}}\right]\Pr(H > 150\,\text{cm}) + E\left[F \mid H = \bar{h}_{<150\,\text{cm}}\right]\Pr(H < 150\,\text{cm})$$

We could continue to break apart the height into ever smaller discrete brackets (h_i), multiplying by the probability of obtaining an individual within that bracket and summing across

$$\hat{F} = \sum_i E\left[F \mid H = \bar{h}_i\right]\Pr(h_i)$$

Ultimately the discrete brackets are reduced to a point where they approximate a smooth continuous distribution. This continuous equivalent is integration and expressed as follows:

$$\hat{F} = \int_i E\left[F \mid H = \bar{h}_i\right]\Pr(h_i)dh_i$$

In words this is explained as the foot size integrated across the distribution of height. It treats height as a nuisance parameter, as we aren't really interested in the height of people but must consider it to calculate the value of interest, in this case foot size.

Each parameter (p) in the analysis can then be treated in this manner to obtain the following:

$$\int_{p_1}\cdots\int_{p_D}\Pr\left(O^1_{1,1}\ldots O^L_{A,R} \mid p_1\ldots p_D\right)\prod_{i=1}^{D}\Pr(p_i \mid p_{i+1}\ldots p_D)dp_1\ldots dp_D$$

where D is the number of parameters (or dimensionality) in the analysis. This rather daunting formula is visually simplified by referring to all parameters as $\boldsymbol{M}$ (for mass parameters):

$$\int_{\boldsymbol{M}}\Pr\left(O^1_{1,1}\ldots O^L_{A,R} \mid \boldsymbol{M}\right)\Pr(\boldsymbol{M})d\boldsymbol{M}$$

We must now also recognize that genotype sets ($\boldsymbol{S}_i$) themselves are treated as nuisance parameters within the *LR* calculation. That is, we don't really care what their value takes; however we must consider them in order to calculate the probability of the evidence. If we add mass parameters into the above equation we obtain the following:

$$\sum_i \Pr(S_i)\int_{\boldsymbol{M}}\Pr(O \mid \boldsymbol{M},\boldsymbol{S}_i)\Pr(\boldsymbol{M})d\boldsymbol{M}$$

In theory it is possible to have different mass parameters for each hypothesis but, since the only factors affecting their values are the genotypes, this is unnecessary. Hence they will cancel out in numerator and denominator of the *LR* so that from the equation above:

$$w_i \propto \int_{\boldsymbol{M}}\Pr\left(\boldsymbol{O} \mid \boldsymbol{M},\boldsymbol{S}_i\right)\Pr(\boldsymbol{M})d\boldsymbol{M}$$

We then obtain a term that looks very much like the denominator and numerator terms of the *LR* at the beginning of this chapter. Note that there are no locus terms in this equation, and this is because it is considering the whole profile at once. There are mathematical advantages to considering the data in this manner, but many disadvantages in the required computer power and comprehensibility of results. Some simplifying assumptions can be made to make the problem tractable:

Assumption 1: Peak heights are assumed to be conditionally independent given S_j and M:

$$\Pr\left(\boldsymbol{O} \mid \boldsymbol{S}_j,\boldsymbol{M}\right) = \Pr\left(\boldsymbol{O}^1_{1,1}\ldots\boldsymbol{O}^L_{AR} \mid \boldsymbol{S}_j,\boldsymbol{M}\right) = \prod_a\prod_r\Pr\left(\boldsymbol{O}^l_{ar} \mid \boldsymbol{S}_j,\boldsymbol{M}\right)$$

And considering the observed data across loci and treating mass parameters as a nuisance the model becomes

$$w_j = \int_{\mathrm{M}} \prod_l \prod_a \prod_r \Pr(O_{ar}^l \mid S_j, \mathrm{M}) \Pr(\mathrm{M}) d\mathrm{M}$$

Assumption 2: The weight across a profile is the product of the weights at each locus

$$w_j = \prod_l w_j^l$$

which gives

$$w_j = \prod_l \int_M \prod_a \prod_r \Pr\left(O_{ar}^l \mid S_j, M\right) \Pr(M) dM$$

In other words the weight of a genotype set at a locus is the probability of obtaining the observed data if that genotype set had given rise to it, integrating across all mass parameters. The full profile weight is the product of the weights of the individual loci.

Even with modern computers it is prohibitively complicated to enumerate this complex multiple integral completely. In order to overcome this limitation the use of methods such as MCMC is employed.

MCMC and Profile Building

DNA profile problems can be highly complex – so complex that even with modern computers it would be impossible to test every possible combination of all parameters. Instead the computer can use a process similar to a game of hot and cold with the DNA profile. This mathematical process is called *MCMC* and allows the computation of complex problems with standard computers.

Figure 9.5 shows a diagrammatic representation of the hot and cold analogy of MCMC. Imagine that each square on the board represents a possible answer to some problem. One possible way to find the best answer would be to start at the top left and work down each possible answer in each row and column and then at the end to choose the answer that gave the

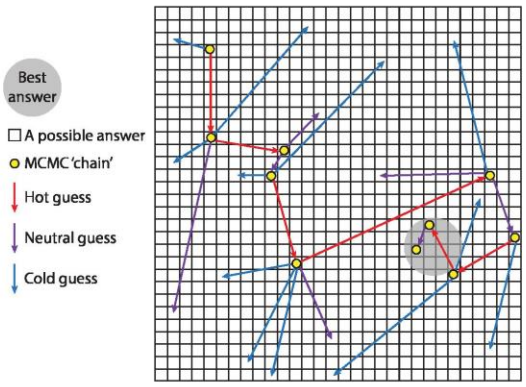


Figure 9.5 Diagrammatic representation of a Markov chain Monte Carlo process as a game of hot and cold.

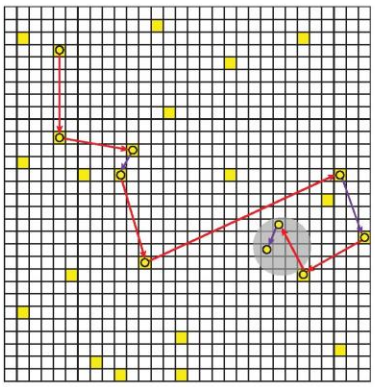


Figure 9.6 Squares landed on (highlighted) in the hot and cold game shown in Figure 9.5.

best solution. Doing this would highlight the solutions that occupy the grey circle on the diagram, but it would take some time as each possible solution would need to be assessed.

Now instead we are going to try and find the good solutions by playing hot and cold. A random point on the board is chosen (in the example below this is near the top left) and the goodness of fit calculated. Next a nearby square is randomly chosen and the goodness of fit calculated. If it is better than the current position it will be adopted and we will 'move' the new position. If the position is much worse than we would be very unlikely to move to it and if it is neutral, then there is a chance we will move there some of the time.

Figure 9.5 shows a series of guesses, some hot, some neutral and some cold, and the eventual path taken by the MCMC chain over the course of 10 moves.

Figure 9.6 shows the answer board, where we have highlighted all the possible solutions that were considered in the journey from starting position to the best answers. It can be seen from Figure 9.6 that only a very small percentage of all possible answers needed to be considered. This is the power of MCMC, its ability to find good answers, or in MCMC parlance 'good sample space', in problems that are very complex, without having to consider every combination of every parameter value in the model.

MCMC Robot

MCMC Robot is available from http://web.uconn.edu/gogarten/bioinf/mcrobot.html or as an application for Apple devices from https://itunes.apple.com/nz/app/mcmc-robot/id454055791?mt=8. This is a useful teaching tool.

Figure 9.7 gives screen grabs from MCMC Robot. The black circles are a probability 'hill'. We can usefully think of this as a contour map. The starting position of the MCMC chain is a blue dot.

In the second screen grab the starting position of the MCMC chain and the first 10 steps are visible.

In the third grab the robot has taken 1000 steps.

In the last grab, in addition to the 1000 accepted steps, the failed steps have been added (in purple).

At the end of these 1000 steps the robot is at the top of the hill and is wandering around. This 1000-step phase is the burn-in. These results are then discarded and the real count begins from the position that was arrived at the end of burn-in. This is to get rid of the 'bad' steps early in the process.

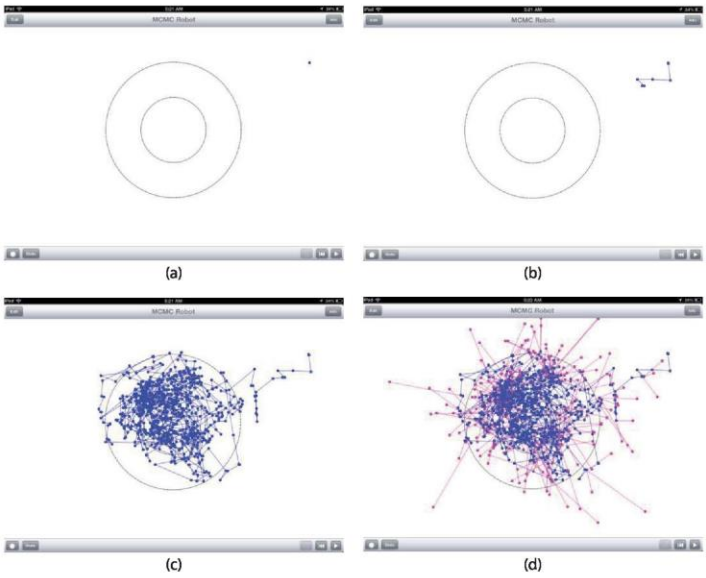


Figure 9.7 Four screen grabs from the MCMC Robot programme. (a) before any steps have been taken, (b) after a small number of steps have been taken, (c) after many steps have been taken and (d) after many steps have been taken and also showing rejected steps.

The Metropolis–Hastings Algorithm

The basis of an MCMC chain is that it steps from one state to another in some sensible way so that it will preferentially sample from the high probability density portion of the sample space. In STRmix the stepping is done using the Metropolis–Hastings algorithm (MHA). MHA compares two states, the current state and the proposed state. The algorithm considers whether to step to the proposed state or stay at the current state. If the proposed state has a higher probability density the chain always steps. If it has a lower probability density it will step some of the time.

It is useful to think of the probability density as a landscape with a hill in it. The objective is to find the top of the hill. The MHA always steps uphill if that is proposed. If the proposal is downhill it does this sometimes. After a while the chain will get to the top of the hill and then wander around sometimes going a way down the slopes.

If the MCMC is sitting at iteration $(y-1)$ and the probability of this current state is $\Pr(y-1)$, the proposed state is y with probability $\Pr(y)$. The MHA will accept the proposed state if $\Pr(y) \geq \Pr(y-1)$ or if a randomly chosen value from $U[0,1] < \dfrac{\Pr(y)}{\Pr(y-1)}$.

Burn-In

When an MCMC starts it is guessing. Mass parameter values are chosen at random and it will be accepting explanations of the observed data that have a very low probability. As the process continues the MCMC will start accepting more and more likely descriptions of the observed data until it has reached an equilibrium state where a small distribution of values for mass parameters and a limited number of genotype sets are being regularly chosen in accordance with how

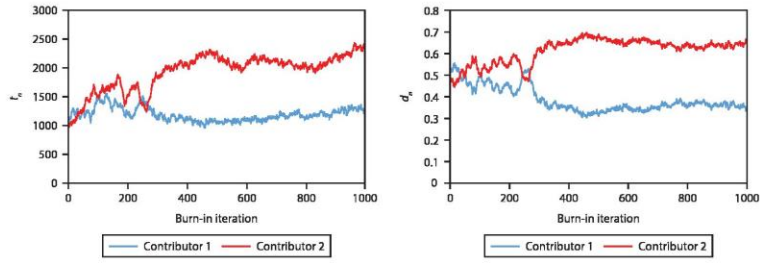


Figure 9.8 Burn-in values for a two person mixture, t_n on left and d_n on right.

well they describe the observed data. The time between when the MCMC starts and when the equilibrium state is reached is termed the *burn-in phase*. In essence it is the time when the MCMC goes from complete randomness to some sensible position.

The first n steps in the MCMC chain are termed *burn-in*, where n is set by the user. These steps are used to get the chain to a reasonable position and the data from these first n steps are discarded. The real count is then started.

The penalty in the first 1/4 of the burn-in for peak heights is $\log(O/E) \sim N(0,0.04)$ and after that it reverts instantly to

$$\log(O/E) \sim N\left(0, \frac{S_a^l c^2}{O_{a+1}^l} + \frac{A_a^l k^2}{E_a^l}\right)$$

The initial constant variance allows the MCMC to find the genotypes more quickly. Figure 9.8 shows two figures giving the t_n and d_n values for a two-person 2:1 mix using a short burn-in of 1000 steps.

Note the resolution of the template amounts for the two contributors. From this output we can tell that the profile must have clearly distinguishable major and minor components. The degradation values for both contributors are very low and do not show any resolution; this is fine. Unlike template amounts the degradation values do not have influence over each other, i.e. just because one degradation value increases does not mean that the other will decrease.

MCMC General

MCMC is a widely used technique outside forensic science and is considered mainstream. It has been used in predicting weather, betting, computational biology, computational linguistics, genetics, code breaking, engineering, physics, aeronautics, stock market and social science.

MCMC is based on a random number generation process. Typically a model will be proposed to describe some data that contains a number of parameters of unknown value. The MCMC trials numerous combinations of parameter values to describe the observed data and ultimately generates posterior distributions for each parameter in the model.

The parameters in the model are the mass parameters (those that are integrated across as nuisance parameters in the previous section). The information supplied by the user is the stutter ratio, the number of contributors and some parameter priors. The data is the observed profile. To picture how the MCMC works for DNA profile interpretation consider the process of building a picture of an expected profile from mass parameters and ultimately comparing it to the observed profile data to calculate a likelihood.

In the following section we show an example of an expected profile being generated from a set of posited parameter values. Note that although the example shows this occurring in a stepwise manner the actual calculations will often occur all together. This means that the order in

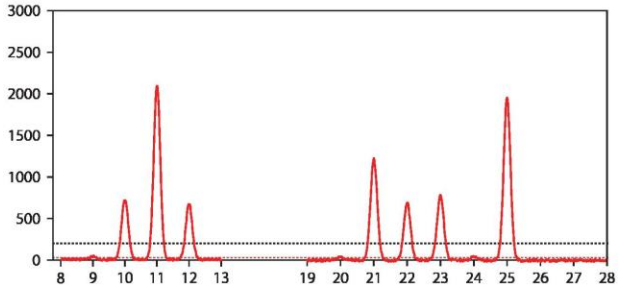


Figure 9.9 An observed two locus profile. Locus 1 has alleles 8 to 13 and locus 2 alleles 19 to 28.

which the various components of the example below occur can easily be changed and would be equally valid.

In Figure 9.9 is an example of an observed two locus profile being analyzed as a two person mixture. Figure 9.10 shows how STRmix builds up an expected profile.

STRmix now has a complete two locus expected DNA profile that it can compare to an observed profile (Figure 9.11).

Converting Mass Parameters to Peak Heights

The peaks in the early stages of Figure 9.10 (before stutter is taken into account) are the total allelic product (T). This is the sum of the allele and the stutter peak. If there are additional replicates they are allowed to scale up or down by a constant factor, R_r. This allows the whole epg to scale but requires the same relative mixture proportions and degradation. The final formula for total allelic product is

$$T^l_{anr} = R_r A^l t_n e^{d_n \times f\left(m^l_a\right)} X^l_{an}$$

The terms R_r, A^l, t_n and d_n are termed *mass variables*, $\boldsymbol{M}$, as they are used to calculate the mass, or total allelic product, of an allele. In the example shown in Figure 9.10 some genotypes were heterozygote and other homozygote. For the homozygote the height is doubled for that individual. This is termed *dose*, X^l_{an}, i.e. the count of allele a at locus l in contributor n. $X^l_{an} = 1$ for a heterozygote with a and $X^l_{an} = 2$ for a homozygote a. The degradation affects peak heights with a dependence on their molecular weight. The term $f\left(m^l_a\right)$ is a function of molecular weight and can be as simple as $f\left(m^l_a\right) = m^l_a$ or can utilize an offset if the desired behaviour of the model is to begin degradation at some minimum molecular weight.

The next step is to introduce stuttering. The back stutter ratio ($N - 1$ repeat), SR, is a function of the sequence (LUS). Hence, different alleles have slightly different stutter ratios. FS is the forward stutter ratio ($N + 1$). T can be apportioned to stutter and allele using the following equations where SR and FS are determined from a model:

$$E^l_{(a-1)} = SR^l_a O^l_a, \qquad E^l_{(a+1)} = FS^l_a O^l_a, \qquad E^l_{an} = \frac{T^l_{an}}{1 + SR^l_a + FS^l_a},$$

where:

$E^l_{(a-1)}$ is the expected back stutter peak height of the ath allele at locus l

$E^l_{(a+1)}$ is the expected forward stutter peak height of the ath allele at locus l

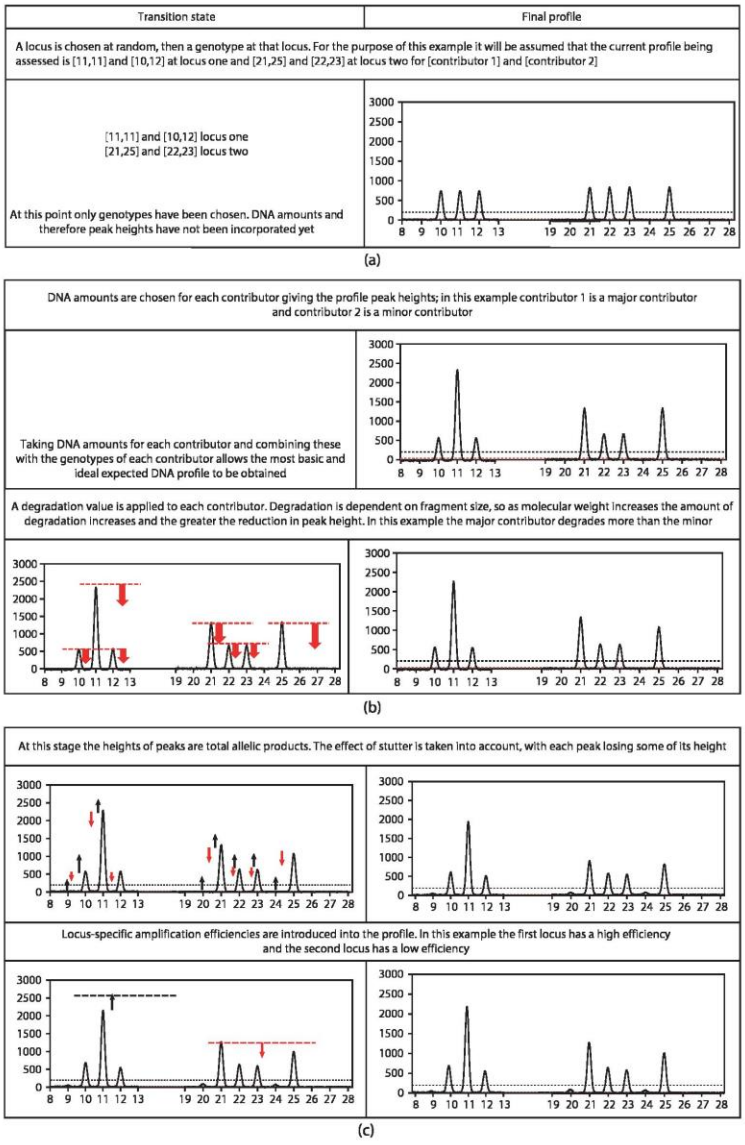


Figure 9.10 Example of building up an expected profile from mass parameters. (a) Choice of genotypes, (b) Development of TAP, and (c) Development of allele and stutter heights.

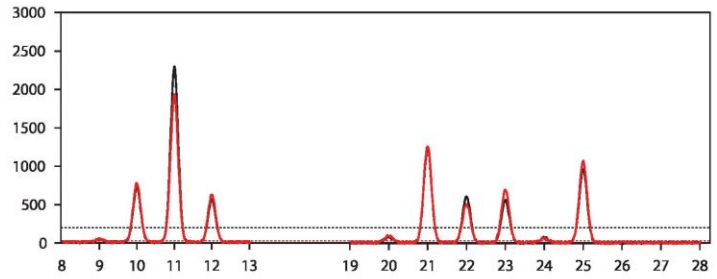


Figure 9.11 Two person mixture expected profile (black) built up from mass parameters overlaying the observed profile (red).

E_{an}^{l} is the expected allelic peak height of the ath allele for the nth contributor at locus l
O_{a}^{l} is the observed height of the ath allele at locus l
T_{an}^{l} is the total allelic product of the ath allele for the nth contributor at locus l
SR_{a}^{l} is the back stutter ratio of the ath allele at locus l
FS_{a}^{l} is the forward stutter ratio of the ath allele at locus l

For the Markov chain we will need $f(O_{ar}^{l} \mid S_q, M)$, which is read as the probability density of the peak at position a given that we know the genotypes and the true template at that allelic position and the one upstream. The modelling above has given the expected heights given that we know the genotypes and the true template at that allelic position. Empirical trials suggest that the relative variance of small peaks is large and that of large peaks is small. The data fit the curve: $\mathrm{var}\left[\log_{10}\left(\frac{O}{E}\right)\right] = \frac{c^2}{E}$, where c^2 is an empirically determined constant. This indicates that the variance of the log of observed over expected heights is inversely proportional to the expected peak height. Most forensic biologists are more familiar with the standard deviation. This is the square root of the variance and is modelled as $\frac{c}{\sqrt{E}}$.

Database Searches Using Continuous Systems

Traditionally, single source profiles or single contributor profiles which have been unambiguously resolved from a mixture have been considered to reach this standard and be suitable for entry into a crime sample database. Single source profiles are relatively simple to interpret, with standard methods generally agreed on and accepted worldwide. Profiles from crime scenes however are frequently compromised in quality. Stochastic events such as heterozygote imbalance, allelic dropout, locus dropout and allelic drop-in can complicate interpretation.[36,37,120] In addition in many cases crime scene samples may be mixed where DNA from more than one individual is present. Stutter, a by-product of the PCR process, can further complicate profile interpretation whenever stutter peaks are of a similar height to the minor allelic peaks in mixed DNA profiles.

Interpretation of these profiles using a continuous model may result in improved profile information and therefore permit database entry. Unless the weight for any given genotype combination is 1, assessing the 'quality' of a profile for its suitability for comparison to a database is not straightforward. A guideline for database entry based on some assessment of the risks of loading an incorrectly inferred profile may be employed where the genotype combination of a contributor is ambiguous, such as $w_i > 0.99$. If an individual's profile cannot be reasonably

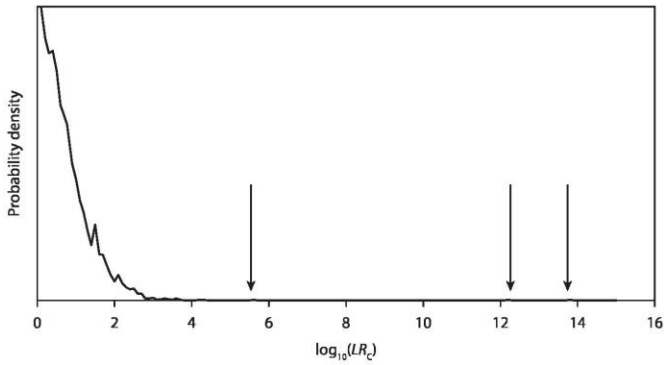


Figure 9.12 Distribution of likelihood ratios (*LR*s) shown on a logarithmic scale, when considering known non-contributing individuals as contributors to a complex three person mixed DNA profile. Only values of $LR > 1$ are shown. Arrows show the *LR* for the three individuals known to make up the mixture. The maximum value for the *LR* when comparing a known non-contributor was 20,348 ($\log_{10}(LR_C) = 4.3$).

inferred from a DNA mixture, regardless of the interpretation method, then it is unsuitable for entry to a database using traditional database methods.

If there is no interpretable single profile from the mixture then a search of the mixture itself should be performed. Comparison of profiles to profiles in a database, where there are multiple possible genotype combinations at one or more loci for matching against known individuals, can be undertaken using the output of a continuous method of interpretation with a modified search algorithm using an *LR* framework.

Each of the individuals on a database can be considered a potential contributor in turn under the following two hypotheses (or others if desired):

H_p: Database individual and $N - 1$ unknown contributors

H_d: N unknown contributors

where N was the number of contributors under consideration. This will provide an *LR* for each individual in the database, compared to the relevant mixture. A cut-off value is then used to reduce the list to a manageable size and remove the most likely potential adventitious matches.

Figure 9.12 shows the results of considering 57,612 individuals as potential contributors to a complex three-person Identifiler® (Thermo Fisher Scientific) profile. Of these 57,612 individuals, 3 were known contributors and 57,609 were known non-contributors.

Of the 57,612 individuals, 4000 gave an *LR* in favour of H_p and as Figure 9.12 shows the majority of these were below an *LR* of 100. In contrast the known contributors gave *LR*s of greater than 400,000 and were clearly distinguishable from the non-contributors.

Continuous *LR*s in Practice

Artificial two and three person mixed DNA profiles with known contributors were amplified with an Applied Biosystems NGM SElect™ multiplex (Invitrogen, Carlsbad, CA) and separated on an Applied Biosystems 3130*xl* capillary electrophoresis instrument.

The *LR* was calculated for each contributor to each of the four two-person and six three-person mixed DNA profiles using both a binary method and the continuous method of interpretation discussed in this paper. The hypotheses considered are as follows:

H_p: The DNA came from P_1 and unknown people up to the number of contributors.

H_d: The DNA came from all unknown people.

LR_B was calculated in MS Excel following the 'F model' as described in Kelly et al.[724]

Table 9.1 shows the *LR* produced from the continuous method as described above and implemented through Java software and a binary method for the same set of mixtures of known source calculated in Excel following the F model described in Kelly et al.[724] Profiles were of reasonable quality to allow assessment by the binary method. Table 9.1 shows the information gain by using a continuous system. In the two person scenarios where the individual profiles can be well resolved the information obtained from the two methods are similar. For three person mixtures, the results of the two systems diverged. LR_B was markedly lower or unable to be determined for three person mixtures, whereas LR_C continued to produce *LR*s consistently much greater than 1.

Table 9.1 shows a mild increase in *LR*s when using the continuous system for simpler two person mixtures, but the real strength comes when three person mixtures are considered. The wastefulness of the binary system is highlighted when complex profiles are analyzed based only on the presence or absence of peaks and do not make use of their height.

There has been a view that peak heights are of limited value at low template. To investigate this concept a cut-down version of STRmix (STRmix™ lite) that ignored heights was tested against the full STRmix version. The full version outperformed STRmix lite in a limited trial both for true and false donors even at very low template (see Figure 9.13 and Table 9.2).

Table 9.1 Likelihood Ratio Results of Continuous vs Binary Method for Assessing Two- and Three-Person Profiles

Continuous			Binary		
Mixed DNA Profiles from Two Contributors					
Person 1	**Person 2**		**Person 1**	**Person 2**	
6.1×10^{15}	3.2×10^{16}		1.2×10^{16}	3.8×10^{15}	
1.9×10^{19}	6.4×10^{19}		9.5×10^{17}	1.9×10^{19}	
2.4×10^{19}	4.9×10^{19}		2.1×10^{18}	9.6×10^{18}	
3.0×10^{17}	1.3×10^{20}		4.0×10^{14}	1.3×10^{20}	
Mixed DNA Profiles from Three Contributors					
Person 1	**Person 2**	**Person 3**	**Person 1**	**Person 2**	**Person 3**
2.7×10^{8}	4.7×10^{11}	3.8×10^{13}	NC	226	1.39×10^{6}
9.1×10^{8}	4.2×10^{11}	2.8×10^{13}	NC	3	14,261
8.4×10^{19}	4.7×10^{11}	6.0×10^{18}	57	<1	331
4.5×10^{12}	5.7×10^{12}	7.20×10^{14}	<1	<1	846
9.6×10^{7}	1.2×10^{20}	1.1×10^{20}	NC	65	317
7.7×10^{18}	3.1×10^{19}	1,254	23	356	NC

NC, non-concordance (and therefore a statistic was not calculated).

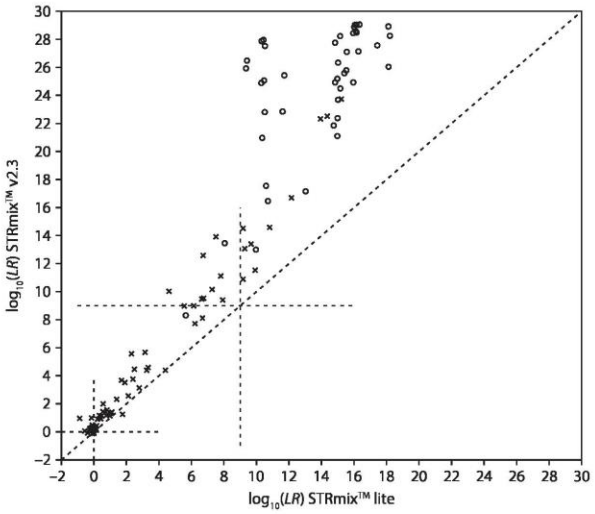


Figure 9.13 $\text{Log}_{10}(LR)$ for STRmix™ lite and STRmix™ v2.3. The diagonal dashed line is $x = y$. Circles represent values that were derived from greater than 50 pg of total input DNA and crosses those less than 50 pg.

Table 9.2 Results of H_d True Tests for a Four-Person 0.25:0.25:0.25:0.25 Mix at 50 pg Total Input Template

		STRmix™ v2.3	STRmix™ Lite
Number of Simulations		12,000,000	10,000,000
H_p True LR		374,104	207
H_d True	p ('1 in')	3,000,000	11,947
	$LR = 0$	99.958%	94.491%
	$LR > 1$	0.0173%	0.0472%
	Average LR	1.005	1.078

LR, likelihood ratio.
Note: Average peak height for the profile was 89 rfu.

In this trial three of the four contributors were input as *knowns*. The *LR*s for both STRmix v2.3 and STRmix lite would be lower if there were fewer knowns. However, for the trial shown the effect is higher *LR*s for true donors and lower ones for false donors.

Comparison with Human Interpretations

A continuous model for DNA interpretations should produce results that are intuitively correct to a trained scientist. We would therefore expect to see a relationship between LR_C and human interpretations.

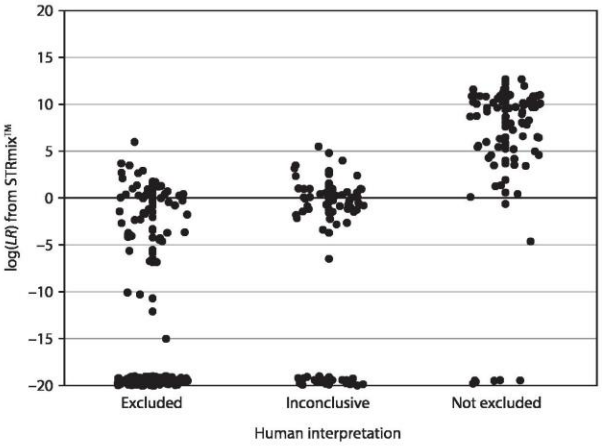


Figure 9.14 Comparison of likelihood ratios (*LR*s) produced using a continuous system with human interpretation. The line at 0 represents neutrality, i.e. the probability of obtaining the profile is the same under propositions of exclusion or inclusion. Results above the 0 line favour inclusion and results below the line favour exclusion. When $LR_C = 0$ (when G_p did not feature in the Markov chain Monte Carlo at any point) the result has been plotted at the bottom of the graph against the *y*-axis label '$LR_C = 0$'.

To test this concept, previously reported casework Profiler Plus® profiles were reanalyzed using the continuous model described. Epgs were analyzed using the continuous model. The samples in this study resulted in 39, 274, 207 and 50 comparisons to single source, two, three and four person mixed profiles, respectively. LR_C produced by the model were compared with the human interpretation for the same result (Figure 9.14). The propositions considered were as follows:

H_p: The DNA came from the POI and unknown people up to the number of contributors.

H_d: The DNA came from all unknown people.

Human interpretations were sorted into three categories: *not excluded*, *inconclusive* and *excluded*.

Inspection of the graph shows a broad alignment of human- and model-based interpretation except that on average human interpretations were more conservative.

Use on Profiles of Varying Quality

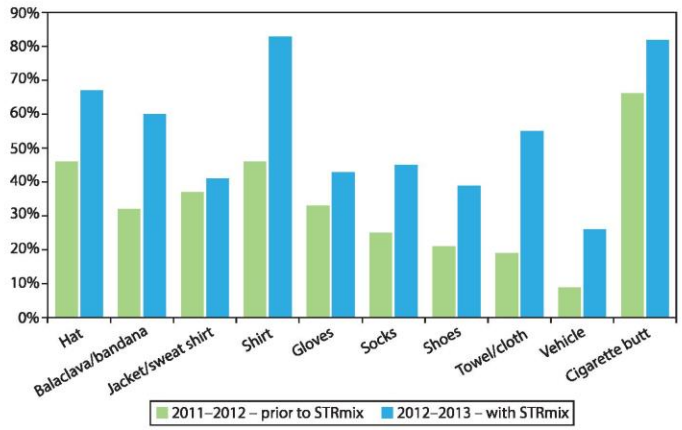
Specificity and sensitivity are not trivial to define when we talk of a software system for DNA profile interpretation. In Ref. 756 (appendix 4) Taylor and Buckleton clone a drop model and show that STRmix extracts useful and correct information from very low level DNA results beyond what would be expected by systems not using peak height (any of the drop models).

To answer the question of sensitivity most directly we suggest that the work described in Ref. 746 and reproduced below in Figures 9.15 through 9.19 might be useful. The *LR* distributions for H_p true and H_d true are very well separated at high template for two person mixtures. As the number of contributors increases and the template lowers the two distributions converge on log(*LR*) = 0. This is the correct result. What it means is that the performance of the software is most dependent on the sample (see Box 9.2). At high template, STRmix correctly and reliably

BOX 9.2 PERFORMANCE OF STRMIX ON VOLUME CRIME

The Institute of Environmental Science and Research Limited (New Zealand) adopted the continuous interpretation software STRmix™ in August 2012. Since then they have been using it for the interpretation of mixtures and the calculation of likelihood ratios for all casework. Within the plot below are the success rates for a number of different sample types submitted for analysis to volume crime team. Success was defined as the ability to obtain a profile suitable for entry to the New Zealand DNA Profile Databank.

The plot shows data for two financial years. The only change in process between the 2011–2012 and 2012–2013 financial years was the introduction of STRmix. The improvement in loading rates is attributable to STRmix's ability to interpret more profiles (data provided by Sarah Scott).



gives a high *LR* for true contributors and a low *LR* for false contributors. At low template or high contributor number, STRmix correctly and reliably reports that the analysis of the sample tends towards uninformative or inconclusive.

Figure 9.20 has been reproduced for Identifiler Plus™ data (courtesy of Erie County Department of Public Safety, Erie County, NY). The laboratory generated 22 four-person mixed DNA profiles. Ten of the profiles were in the approximate proportions of 4:3:2:1. The amount of DNA corresponding to the smallest contributor ranged from 100 pg to 0.625 pg. Twelve of the profiles were in equal proportions (1:1:1:1), where the amount of DNA from each contributor ranged from 400 pg to 1.25 pg. Three and two contributor profiles were prepared similarly. These profiles represent the 'worst' types of profiles likely to be encountered by the laboratory. Each profile was interpreted in STRmix and compared to the four known contributors and 200 known non-contributors. The non-contributors were generated artificially using a Caucasian allele frequency database.[502] A plot of the log(*LR*) versus DNA per contributor (pg) for each dataset is provided in Figure 9.20. The per-contributor amount for H_d true contributors was taken as the average of the known contributors.

Inspection of Figure 9.20 shows the separation of the *LR* distributions for H_p and H_d true propositions is less for Identifiler Plus profiles than for GlobalFiler profiles. Identifiler Plus has six fewer loci used within the *LR* calculation which explains the lower discrimination.

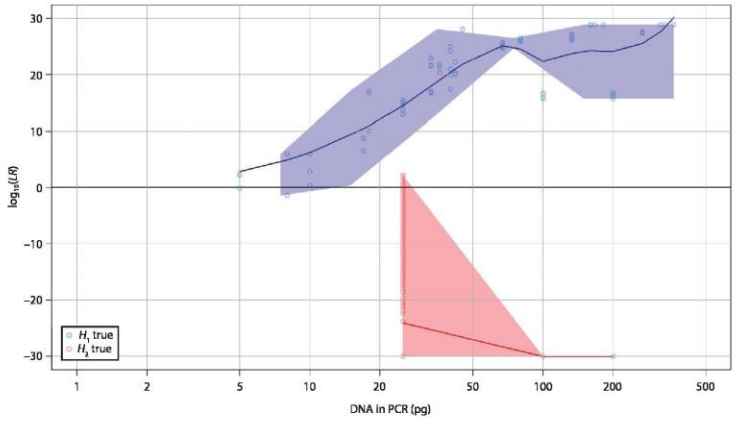


Figure 9.15 Likelihood ratios produced for two person mixtures, with LOWESS lines and polygons showing coverage of scatterplot points.

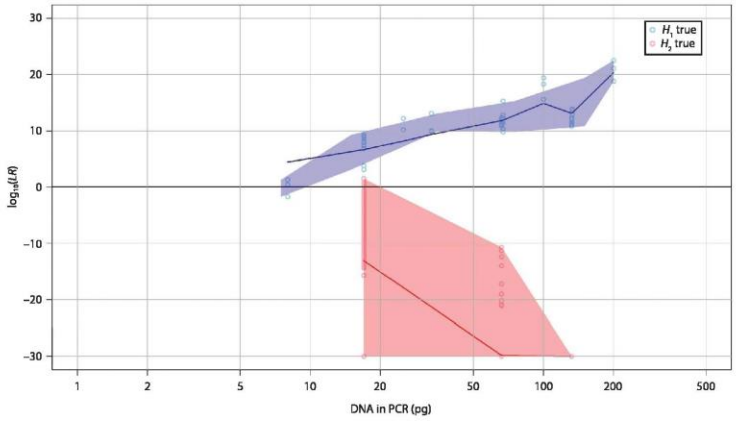


Figure 9.16 Likelihood ratios produced for three person mixtures, with LOWESS lines and polygons showing coverage of scatterplot points.

Results from Bille et al.[699] are given in Figure 9.21 below for a number of trials on true donors. The broad conclusion is that a continuous approach gives better performance for true donors across a range of mixture ratios and template.

Diagnostics for Continuous Systems

Ground truth comparisons should produce a large *LR* when the prosecution proposition is true and a low one when the defence proposition is true. Any results in the opposite direction should have a detectable cause, such as very poor PCR amplification of the known contributors.

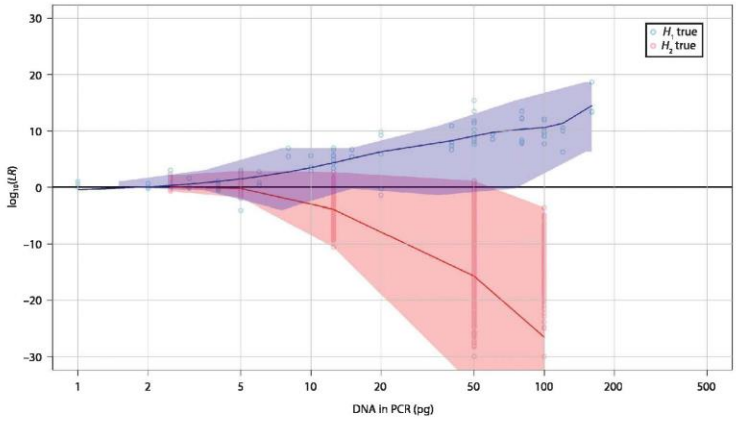


Figure 9.17 Likelihood ratios produced for four person mixtures, with LOWESS lines and polygons showing coverage of scatterplot points.

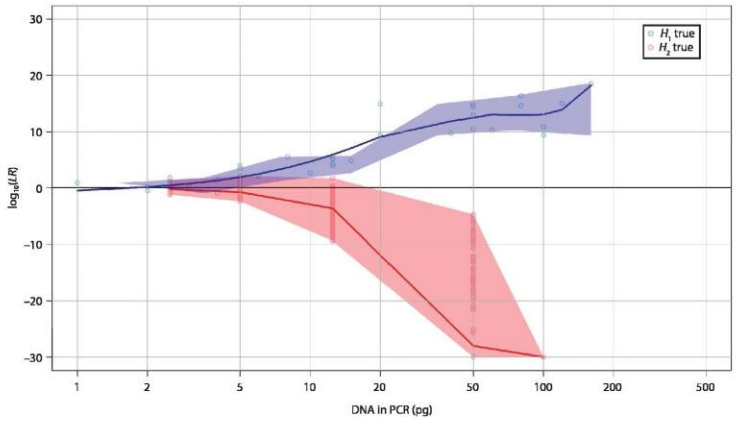


Figure 9.18 Likelihood ratios produced for four person mixtures using three replicate amplifications, with LOWESS lines and polygons showing coverage of scatterplot points.

Equally, occasionally a false contributor may give a high *LR*. This is termed an *adventitious match*. Such tests give little guidance as to whether the *LR* is too large or not large enough but sensible limits may be placed. For example, a two person mix cannot exceed the expected result that would have occurred if it was fully resolvable.

A large number of mixtures where the ground truth is known have been run in STRmix and published in peer-reviewed journals. H_p true trials are comparisons to the known contributor to a profile. H_d true trials are comparisons to known *non-contributors* (i.e. individuals who have

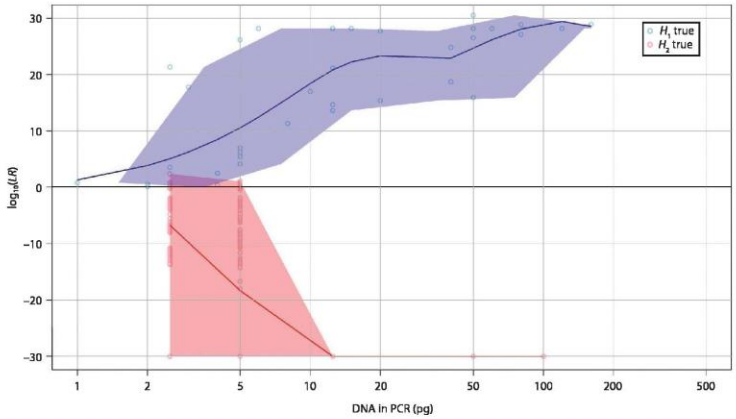


Figure 9.19 Likelihood ratios produced for four person mixtures using three replicate amplifications and assuming three out of the four known contributors in each analysis, with LOWESS lines and polygons showing coverage of scatterplot points.

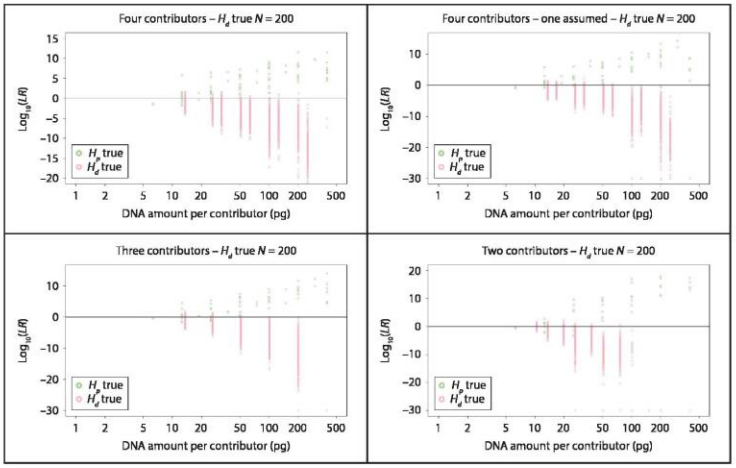


Figure 9.20 Likelihood ratios produced for four, three and two person mixtures from Erie County, PA, NY.

not contributed DNA to the profile). We expect high *LR*s for the true contributors and low *LR*s for the false ones.[746,757]

Turin showed that the average *LR* for the H_d true tests should be 1 (quoted in Good[250]).

Following from this we can state: 'The probability (p) of observing a likelihood ratio of LR_{POI} or larger from an unrelated non-donor is less than 1 in LR_{POI}'.

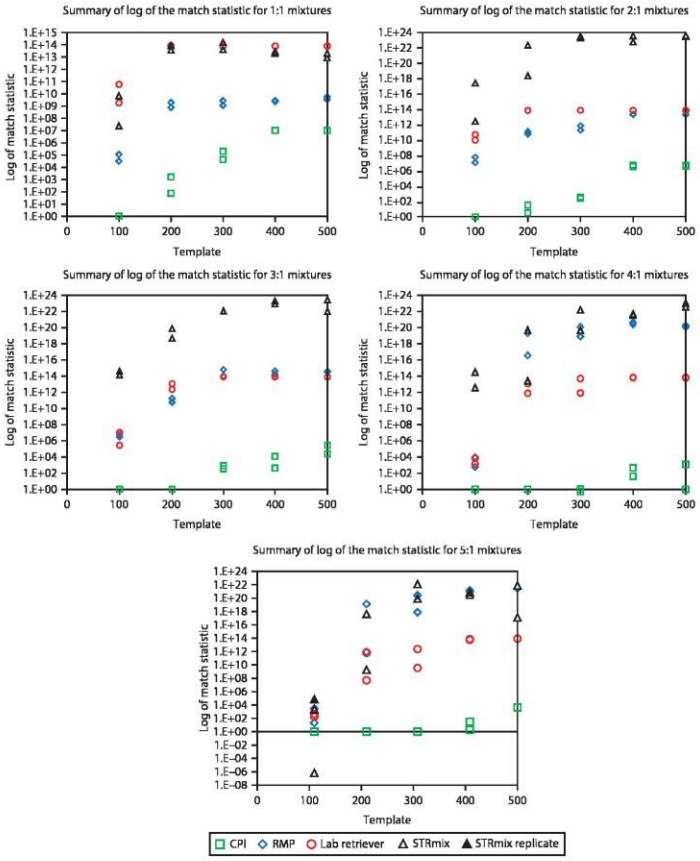


Figure 9.21 A summary of some comparisons. (From Bille, T.W., et al.: Comparison of the performance of different models for the interpretation of low level mixed DNA profiles. *Electrophoresis*. 2014. 35(21–22). 3125–3133. Copyright Wiley-VCH Verlag GmbH & Co. KGaA. Reproduced with permission.)

This gives a frequentist-sounding interpretation to the *LR* but is actually a statement that follows from the laws of probability.

Below we reproduce some examples of these principles generated from in vitro profiles and STRmix. The false donors were simulated using the product rule.

Other published ground truth known tests include the following:

1. GlobalFiler: 264 H_p true; 17,406 H_d true[746]
2. Identifiler: 57 H_p true; 1,902,524 H_d true[757]
3. Identifiler: 54 H_p true; 54,000 H_d true[758]
4. Varying multiplexes: 21 H_p true; 123,230,000 H_d true[759] (results appear in Table 9.3)

Table 9.3 Results of Comparisons of Simulated Random References

Experiment	Number of Contributors	Simulations	Number of H_d True with $LR = 0$	Average H_d True LR	H_p True LR(s)	p, '1 in'
1	4	12,000,000	11,994,959	1.0046	374,104	3,000,000
2	4	10,000	0	0.977	9	44
3	4	120,000	0	0.927	4	29
					7	56
					5	34
					6	49
4	1	80,000,000	79,998,779	1.001	312,325	6,666,666
5	1	100,000	99,618	1.022	215	262
6	2	10,000,000	9,898,155	1.017	278	4,163
					12,557	78,125
7	3	1,000,000	922,585	0.906	234,738	>1,000,000
					2,530	17,241
					43	2,262
8	1	10,000,000	9,999,960	0.872	218,070	250,000
9	1	10,000,000	9,274,620	1.003	14	14

LR, likelihood ratio.
Note: Multiple results, where shown, are due to multiple unknown contributors under H_d.

Gelman–Rubin Convergence Diagnostic $\hat{R}$

This statistic gives an indication of whether the chains have converged. It compares the within-chain variance (W) and between-chain variance (B) for M chains each of n measurements. This means that more than one chain must be run in order to use this method. To visualize the effect imagine that the chains have each chosen one corner of the space. Then the between-chain variance might be high and the within-chain low. This is a symptom of non-convergence. If the chains are all intertwined across similar space then the within- and between-chain variances are similar:

$$\hat{R} = \sqrt{1 - \frac{1}{n}\left(1 - \frac{B}{W}\right)}$$

If we set the between-chain variance, B, approximately equal to the within-chain variance, W, we can see that $\hat{R}$ tends to 1. If $\hat{R} > 1.2$ (approximately) then the chains may not have properly converged. In practice, the performance of the Gelman–Rubin Convergence Diagnostic is not to the expected level.

Effective Sample Size

The successive states from an MCMC chain are correlated.

If you were to look at each set of adjacent points compared to the mean, they would be correlated (both above or both below): A, both *a*bove; B, both *b*elow; C, changed.

If the points in Figure 9.22 were independent then we would expect equal numbers of As and Bs and twice as many Cs. If they are correlated then As and Bs will outnumber Cs. Figure 9.22 shows 15 As, 19 Bs and 5 Cs; therefore the points in this MCMC separated by one iteration are correlated.

The next step tries points that are separated by two iterations, as shown in Figure 9.23.

Figure 9.23 shows 6 As, 8 Bs and 5 Cs and the data are therefore still correlated. This process continues until the correlation is 0. This is called the *correlation at lag* k. Figure 9.24 shows the correlation for a real STRmix calculation at lag 0 to 25,000. In this example the correlation is not 0 until examining values are separated by approximately 8000 iterations.

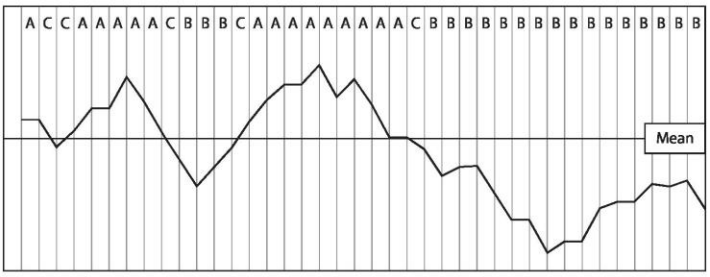


Figure 9.22 Tracking of a Markov chain Monte Carlo property over a number of iterations.

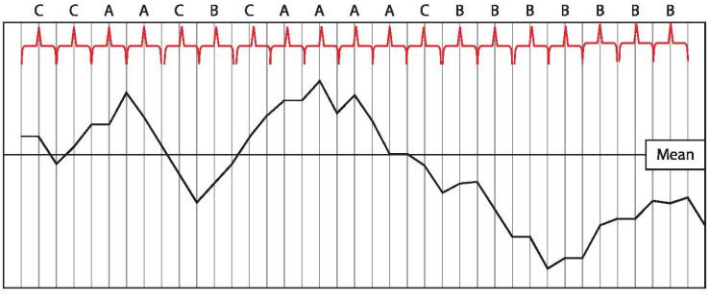


Figure 9.23 Tracking of a Markov chain Monte Carlo property over a number of iterations, looking at pairs of values separated by two iterations.

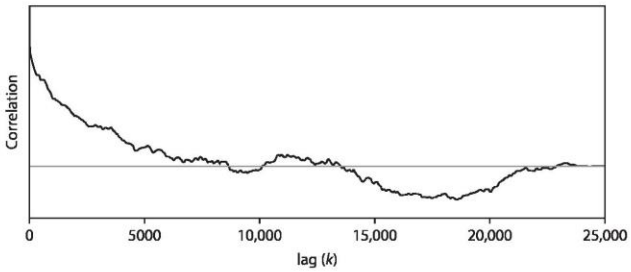


Figure 9.24 Correlation at lag *k* for an STRmix™ calculation.

Once we know how correlated the data are, we can determine how many *independent* samples the MCMC has run for. This is called the *effective sample size* (ESS) of the MCMC. We can then use the effective sample size and the weights to calculate the effective count, i.e. the number of independent counts that genotype set S_j was the focus of the MCMC.

Getting the Number of Contributors Wrong

STRmix requires the assignment of a number of contributors (hereafter N) that have donated to those alleles showing peaks above the selected analytical threshold (AT). However N is never known with certainty. It may be useful to start the thought process from first principles.

$$LR = \frac{\sum_{n} \Pr(E \mid N = n, H_p) \Pr(N = n \mid H_p)}{\sum_{m} \Pr(E \mid N = m, H_d) \Pr(N = m \mid H_d)} \tag{9.1}$$

We remind ourselves that what is behind the bar is assumed to be true and what is in front of the bar is unknown. This informs us that we do not determine N from E, the evidence. The rationale for our current process is that we assume $\Pr(N = n|H_p) = \Pr(N = m|H_d) =$ constant for all n and m. This assumption may be motivated by the observation that the information in H_p and H_d seldom informs N strongly. Hence

$$LR = \frac{\sum_{n} \Pr(E \mid N = n, H_p)}{\sum_{m} \Pr(E \mid N = m, H_d)} \tag{9.2}$$

There is no need for the distributions of m and n to be the same in the numerator and denominator. However we are cognizant both of the reality of and perception of bias and of fitting the profile to the POI. So, initially at least, we will constrain n to equal m. This is also a current technical limitation of STRmix up to v2.3.

$$LR = \frac{\sum_{n} \Pr(E \mid N = n, H_p)}{\sum_{n} \Pr(E \mid N = n, H_d)} \tag{9.3}$$

The summation in the denominator is usually dominated by one term. This term is when n is the number that best explains the profile. Given the constraints above, this should also return on average not only the best approximation to Equation 9.2 but one that is fair and reasonable to the defence.

Assigning N. Recall before we start that N is unknown and unknowable. We have plausibly added to confusion by suggesting that N should be determined. A much better word would be *assigned*. In most cases an N that optimizes $\Pr(E|N = n,H_d)$ is obvious. It is only rarely to the advantage of the defence to posit an additional unknown beyond that required to explain the profile. These situations tend to occur for high order mixtures and are discussed later. Recall that we seek to maximize $\Pr(E|N = n,H_d)$, not minimize the *LR*. The *LR* can always be minimized to at least 1 if that was the goal.

For STRmix, E is composed of those peaks above AT. Therefore we seek a reasonable maximization of $\Pr(E|N=n,H_d)$ where E is confined to those peaks above AT. This is usually obtained by the minimum N that is needed to explain E and would usefully consider peak heights and balances. We will term this assigned number N_{Hd}^{E}.

What Happens If We 'Add One'?

In our experience if N is set to one larger than N_{Hd}^{E} one of a number of things happens:

1. STRmix splits the smallest contributor.
2. STRmix adds a trace that is scattered widely among genotypes including dropped alleles.

Behaviour 1 tends to happen if there is no evidence in E to suggest another contributor (termed *situation 1*). Behaviour 2 tends to happen if there is some evidence in E to suggest a trace contributor (termed *situation 2*). However we are unable to guarantee that this list is completely exhaustive.

Where uncertainty exists in the number of contributors a routine policy of the 'addition of one' is not recommended. It is advised that replication by PCR of the profile is attempted to help inform the decision. The addition of one when there is little or no evidence in E to do so increases the risk of adventitious hits.[757,758] This effect is highly undesirable.

What about Sub-Threshold Peaks?

Imagine that situation 1 holds but that there are considerable indications of an additional contributor below AT. We will term this E_{sub}. First it is essential to note that modern multiplexes show a range of artefacts, including but not limited to forward stutter, double backward stutter and for SE33 –2bp stutter. All of these may appear in the sub-threshold region if a large allelic peak is present.

Four potential policies come to mind:

1. Rework (re-PCR) the profile in order to help confirm the number of contributors.
2. 'Add one'.
3. Lower the AT.
4. Deem the profile uninterpretable after checking for exclusionary potential.

In all cases where sub-threshold peaks suggest uncertainty in the number of contributors, replication should be the first option. Of the other options STRmix is likely to cope best with the third policy. This will allow STRmix to 'see' the same information as the operator is using.

To implement this approach it is necessary that validation and especially investigation of the drop-in parameters have been done to at least as low as the AT will be lowered.

The risks of the 'add one' option were discussed above.

When Is $\Pr(E|N=n,H_d)$ Maximized?

It is relatively easy to find the N that maximizes $\Pr(E|N=n,H_d)$ if peak heights are ignored; however this is not useful for continuous systems that use height information. STRmix itself does not provide access to this information and minimizing the LR across N does not achieve this.

We can construct profiles where an $N=4$ mixture looks exactly like an $N=3$ mixture and NIST 13 case 5 is one of these.[760] In such a case it is the lower N not the higher that optimizes $\Pr(E|N=n,H_d)$. The choice of an N that is lower than the true (but unknown) runs an increased risk of false exclusion, not false inclusion.

We do not have experience of a situation where an N greater than that needed would optimize $\Pr(E|N=n,H_d)$; however it is not clear to us how we would have known that such a situation had occurred. It seems likely to us that the safest policy is to set N to the lowest number

that effectively explains the profile when considering peak heights. We accept the subjectivity inherent in this statement. It is possible that for very high order mixtures this assessment is very difficult. These profiles may be uninterpretable with current technology.

What If the POI Does Not Fit for the Assigned *N* under Pr($E|N = n, H_d$)?

It is possible that under N^E_{Hd} the POI is excluded but under $N^E_{Hd}+1$ he or she is not. To examine this it is helpful to go back to Equation 9.2:

$$LR = \frac{\sum_n \Pr(E \mid N = n, H_p)}{\sum_m \Pr(E \mid N = m, H_d)}$$

In the case described it is likely that the best approximation to Equation 9.2 is achieved by

$$LR = \frac{\Pr(E \mid N = N^E_{H_d}+1, H_p)}{\Pr(E \mid N = N^E_{H_d}, H_d)}$$

Versions of STRmix up to v2.3 cannot implement this and it is unlikely that

$$LR = \frac{\Pr(E \mid N = N^E_{H_d}+1, H_p)}{\Pr(E \mid N = N^E_{H_d}+1, H_d)}$$

is a fair and reasonable assessment.

Empirical Trials

The effect of the uncertainty in the number of contributors has been reported for a number of profiles with N and $N + 1$ assumed contributors, where N is the known number of contributors.[757,758] The inclusion of an additional contributor beyond that present in the profile most often had the effect of lowering the *LR* for trace contributors within the profile. STRmix most often adds the additional (unseen) profile at trace levels which interacts with the known trace contribution, diffusing the genotype weights and lowering the *LR*. There was no significant effect on the *LR* of the major or minor contributor within the profiles.

Addition of One Contributor

A selection of one, two and three person mixtures was interpreted as two, three and four person profiles, respectively. The *LR* was calculated for both the known contributors and 200 known non-contributors. A summary of the *LR*s assuming the correct and one additional contributor is given in Table 9.4.

Most of the time the *LR*s for the known contributors were affected very little. The four largest changes downwards are shown in bold. This means that the wrong assumption leads to a lower *LR*. The five largest changes upwards are shown in red. In these cases the wrong assumption leads to a larger *LR*.

For the 200 or more known non-contributors the distribution of *LR*s is given in Figures 9.25 through 9.27.

Subtraction of One Contributor

A two contributor profile was adjusted by artificially adding a third contributor. The third contributor was constructed as a child of the two known contributors and therefore shared alleles at all loci. In this way it was possible to confuse the true number of contributors. The child was added in the varying average heights 50 rfu, 100 rfu and 200 rfu. At higher amounts the evidence of a third contributor would be clear. Each artificially constructed three person profile

Table 9.4 Summary of the Likelihood Ratio for Profiles Assuming the Correct and One Additional Contributor

Known Ground Truth	Likelihood Ratio			
	Assumed Number of Contributors			
	1	2	3	4
Single-source samples	4.19×10^{20}	4.19×10^{20}		
	4.19×10^{20}	4.19×10^{20}		
	2.65×10^{23}	2.65×10^{23}		
	2.65×10^{23}	2.65×10^{23}		
	6.81×10^{19}	6.81×10^{19}		
Two-person mixtures		6.31×10^{17}	5.01×10^{17}	
		2.51×10^{16}	2.51×10^{16}	
		6.31×10^{19}	7.94×10^{19}	
		2.51×10^{16}	2.51×10^{16}	
		3.98×10^{22}	3.98×10^{22}	
		2.00×10^{15}	1.58×10^{12}	
		3.98×10^{22}	3.98×10^{22}	
		7.94×10^{14}	3.98×10^{14}	
		3.98×10^{22}	3.98×10^{22}	
		2.51×10^{9}	1.00×10^{9}	
		3.98×10^{22}	3.98×10^{22}	
		3.16×10^{9}	7.94×10^{6}	
		3.98×10^{22}	3.98×10^{22}	
		7.94×10^{15}	6.31×10^{15}	
		3.98×10^{22}	3.98×10^{22}	
		6.31×10^{15}	5.01×10^{15}	
Three-person mixtures			9.77×10^{12}	1.05×10^{13}
			3.09×10^{21}	2.75×10^{21}
			5.37×10^{12}	5.89×10^{12}
			4.79×10^{12}	2.24×10^{13}
			1.86×10^{21}	2.09×10^{21}
			5.01×10^{12}	2.40×10^{13}
			2.75×10^{3}	2.24×10^{3}
			2.34×10^{24}	2.34×10^{24}
			1.86×10^{3}	7.24×10^{2}
			26.3	15.8

(Continued)

Table 9.4 *(Continued)* Summary of the Likelihood Ratio for Profiles Assuming the Correct and One Additional Contributor

Known Ground Truth	Likelihood Ratio			
	Assumed Number of Contributors			
	1	2	3	4
			2.34×10^{24}	2.34×10^{24}
			6.17	4.57
			2.69×10^{3}	5.25×10^{3}
			2.34×10^{24}	2.34×10^{24}
			7.94×10^{10}	4.37×10^{9}
			4.79×10^{9}	1.12×10^{10}
			3.63×10^{13}	1.17×10^{15}
			6.76×10^{11}	5.13×10^{12}
			4.79×10^{9}	5.62×10^{9}
			6.61×10^{12}	2.51×10^{13}
			1.74×10^{9}	4.07×10^{9}
			5.13×10^{16}	**3.09×10^{12}**
			3.09×10^{13}	**9.12×10^{10}**
			9.77×10^{8}	**9.33×10^{7}**
			2.51×10^{20}	2.51×10^{20}
			2.57×10^{14}	**3.55×10^{12}**
			24.5	60.3

Note: Bold text represents the four largest downward changes; red text represents the five largest upward changes.

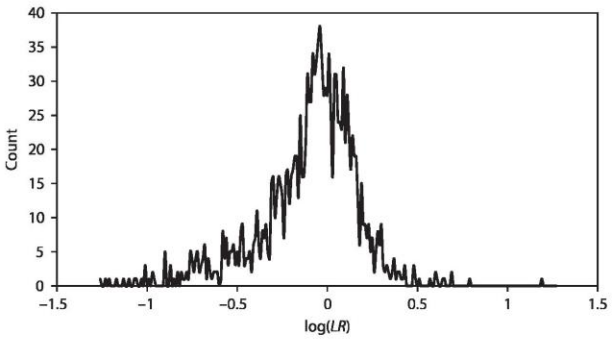


Figure 9.25 Distribution of adventitious link likelihood ratios for single source profiles interpreted as two person mixtures.

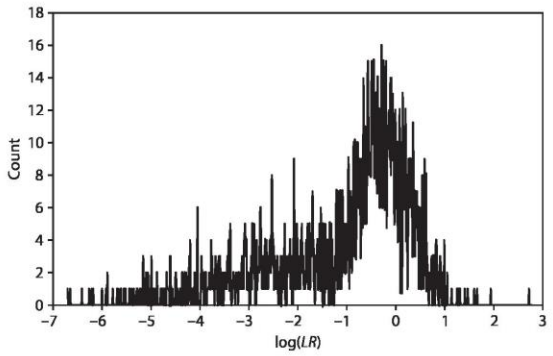


Figure 9.26 Distribution of adventitious link likelihood ratios for two person mixtures interpreted as three person mixtures.

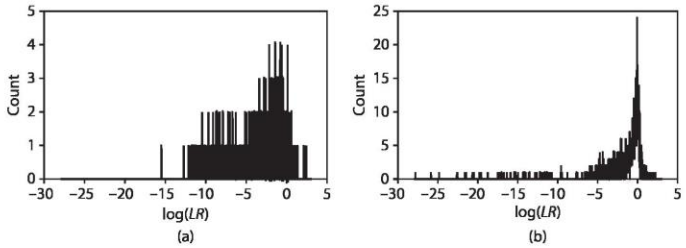


Figure 9.27 Distribution of adventitious link likelihood ratios for three person mixtures assuming three (a) and four (b) contributors.

was interpreted assuming two contributors and compared with the three known contributors and 200 non-contributors.

The *LR* of the two contributors was not affected in any of the trials. For all trials 216 known non-donors all returned $LR = 0$.

TH01 9.3,10

In some circumstances where 9.3 and 10 at TH01 are present in the same mixture one can end up as an unresolved shoulder on the other. In some circumstances this will cause a false exclusion. The diagnostic is $LR_{TH01} = 0$ and $LR > 1$ elsewhere. The recommended action is to check for exclusionary potential at TH01 and if none exists to exclude the locus. We have no instances of it occurring but by analogy this problem is likely for all 0.1 and 0.3 variants at tetra allelic repeat loci, 0.1 and 0.4 at penta allelic repeat loci, and 0.1 and 0.2 at triallelic repeat loci.

Very Significantly Overloaded Samples

STRmix has a function for overloaded profiles. This function allows the examination of some overloaded profiles. However in some circumstances false exclusions eventuate. This has occurred for very significantly overloaded profiles of the order of 3 ng amplified. It is highly preferable to avoid significantly overloaded profiles.

Triallelic Loci

STRmix has no modelling currently for triallelic loci. If a reference is triallelic (either type I or II) and is present in significant proportion, a false exclusion is possible. The diagnostic is $LR = 0$ for the triallelic locus and $LR > 1$ elsewhere. The recommended action is to check for exclusionary potential at the triallelic locus and if none exists to exclude the locus. The action should appear in the report.

Use and Acceptance of Continuous Systems in Courts

The Frye standard[761] arises from the case *Frye v United States*, 293 F. 1013 (D.C. Cir. 1923) in which the court gave the following opinion:

> Just when a scientific principle or discovery crosses the line between the experimental and demonstrable stages is difficult to define. Somewhere in this twilight zone the evidential force of the principle must be recognized, and while the courts will go a long way in admitting experimental testimony deduced from a well-recognized scientific principle or discovery, the thing from which the deduction is made *must be sufficiently established to have gained general acceptance* in the particular field in which it belongs.

This passage emphasizes that the deduction must proceed from a well-recognized scientific principle or discovery. Moving to software this would appear to mean that the software must implement accepted scientific principles. We would not read this as meaning that the software itself must be in prevalent use but that the principles upon which it is based must be generally accepted. This is thoroughly sensible. Obviously when any software first appears it will be in limited use but it may be very soundly programmed from well-accepted principles. The court clearly envisages that the standard is that the principles are sound, not some sort of vote about how often the software is used.

Any developer should outline the principles upon which the software is based and ensure that these meet the standard. We outline these for the STRmix software in Table 9.5.

Agreement in science proceeds by the peer-reviewed literature. We have surveyed the literature using the Scopus online search tool. We searched for the keywords 'forensic and DNA and interpretation' for the time period from 2012 to January 2015 and obtained 150 references. These were scored as being for the use of probabilistic genotyping, against or irrelevant. We obtained 39 references for, 1 against[764] and 110 that were not relevant. A key phrase from the one scored as against probabilistic genotyping was, "The variance of heterozygote balance was more expanded in two person mixtures than in one-person samples. Therefore it is not suitable to use allelic peak heights/areas for estimating the genotypes of the contributors such as the quantitative analysis'.

Table 9.5 Evidence of Acceptance for Some Principles Underlying the STRmix™ Software

Principle	Evidence of Acceptance
Markov chain Monte Carlo	This is very standard mathematics employed in many areas of science. Searching the term *Markov chain Monte Carlo* in Scopus returns more than 22,000 records. Scopus is an online bibliographic database containing abstracts and citations for 20,000 peer-reviewed academic journal articles.
Stutter and peak heights and the variance about them can be predicted from empirical models	Studies of stutter and allele peak heights are now quite numerous and have appeared in the peer-reviewed literature.[120,762,763]
The probability of a multilocus genotype can be estimated from allele probabilities and the coancestry coefficient	For STRmix the model follows Balding and Nichols.[358] This model is based on published literature and appears as NRC II Recommendation 4.2. It is the most conservative of the methods in common forensic use.[365]

The right to confront adverse witnesses is ancient. It appears in the Acts of the Apostles 25:16, when Roman governor Porcius Festus states when discussing the proper treatment of Paul: 'I answered them that it is not the custom of the Romans to hand over any man before the accused meets his accusers face to face, and has an opportunity to make his defense against the charges' (New American Standard 1977). It is also cited in Shakespeare's *Richard II*: 'Then call them to our presence; face to face, And frowning brow to brow, ourselves will hear The accuser and the accused freely speak'.

The European Court of Human Rights, Article 6(3), provides that 'everyone charged with a criminal offence' has the right to 'examine or have examined witnesses against him'. This basically means that the accused, or his or her lawyer, should have a chance to put questions to adverse witnesses. The Sixth Amendment to the Constitution of the United States provides that a person accused of a crime has the right to confront a witness against him or her in a criminal action. This includes the right to be present at the trial as well as the right to cross-examine the prosecution's witnesses. It is therefore essential that the witness can represent the evidence and meet the needs of cross-examination. No analyst can be expected to understand the mathematics and computer programme to the extent that they could recreate the system, except the developers themselves. However it is an expectation that analysts at least understand the workings of any system they use to be able to understand and explain the results.

In *R v Noll* (*R v Noll* [1999] 3 VR 704) the witness acknowledged that although his evidence was based on accepted scientific theory he himself could not describe that theory. At appeal it was submitted that this meant the witness was incapable of giving the DNA evidence and should have been excluded. The court found that although the witness was unable to explain the technical aspects of the theory, he was entitled to rely on other expert opinion. Addressing this issue specifically Justice Ormiston explains:

> Professional people in the guise of experts can no longer be polymaths; they must, in this modern era, rely on others to provide much of their acquired expertise. Their particular talent is that they know where to go to acquire that knowledge in a reliable form.

It has yet to be established in either the scientific or legal frames what level of comprehension is required of a witness giving testimony based on probabilistic methods. However it is clear that this must at least meet minimum levels that ensure that inappropriate testimony is not given. In this document we will attempt to progress the establishment of such standards.

Of note, the International Society for Forensic Genetics (ISFG) *2012 Guidelines* recommend probabilistic methods. In the United States, both the Scientific Working Group on DNA Analysis Methods and the Organization of Scientific Area Committees DNA Analysis 2 subcommittee are working on guidelines for assessing probabilistic genotyping software tools. In addition a considerable number of modern probabilistic genotyping software programmes have been or are being developed by researchers or academics, often with very strong mathematical or statistical backgrounds.[107,108,110,111,686,693,725,742,743,747]

Open Source

Open source software (OSS) is computer software with its source code made available with a license in which the copyright holder provides the rights to study, change and distribute the software to anyone and for any purpose.*

In the forensic DNA field several programmes involved in the interpretation of DNA are open source. This is supported by the ISFG. There are however clear pros and cons to this approach. The goal of OSS was, in part, to invite collaborative development. We are aware of only one attempt at this and that was the development of Lab Retriever[108] from LikeLTD.[109,114,462]

* Wikipedia.

This initiative did not meaningfully improve LikeLTD except in the area of interface at the cost of introducing two new errors into the software.*

Both Lab Retriever and LikeLTD maintain a variant population genetic model regrettably introduced by Balding and Buckleton.[114] We lament the addition of another variant in the forensic field as it further fragments practice.

Balding reports that the reprogramming of LikeLTD into Lab Retriever did result in a number of minor bugs being noticed in the original software and this is clearly beneficial. Only one of these affected the math. Balding reports that it did not affect any casework. However with open source it is difficult to know exactly who has used the software in which cases and therefore hard to inform them of a bug in previously used versions. LikeLTD maintains a mailing list and Lab Retriever reports bugs on their website.

Collaborative development is easy to achieve without recourse to OSS. In the STRmix instance Professor James Curran has proposed many useful additions or amendments to the code that have improved runtime performance greatly. In fact a great many scientific collaborations proceed every day without the use of open source.

Probably the primary perceived benefit was openness. The placement of the code in the public domain allowed open scrutiny of the code. In our own work we have never discovered a bug by examination of the code. This has always occurred either by examination of intermediate results in the process or observation of an aberrant result in testing. The two bugs described for Lab Retriever were discovered by the authors, not from the code but by repeating simple calculations.

There have been applications, unsuccessful to date, by defence to access code from closed source software.† A summary of known applications is provided below.

1. A request for the source code of the Office of the Chief Medical Examiner's (OCME, New York) Forensic Statistical Tool (FST) was denied by New York County Supreme Court Justice Carruthers in his May 2, 2012 decision in *People v William Rodriquez*, Indictment No. 5471/2009. The court declined to sign a judicial *subpoena duces tecum* compelling the OCME to produce the source code of the FST.

2. Kings County Supreme Court Justice Dwyer similarly denied a request by defence to compel the OCME to disclose FST's source code in *People v Collins*, Indictment No. 8077/2010 and *People v Peaks*, Indictment No. 7689/2010.

3. In *Commonwealth of Virginia v Brady*, an oral decision on July 26, 2013, Honourable W. Allan Sharrett denied the request for True Allele's source code, ruling that 'validation studies are the best tests of the reliability of source codes. In this case validation studies have been performed, and they have been performed with positive results. [The validation studies] have shown, in fact, that it has been proven to be reliable'.

4. In *Commonwealth of Pennsylvania v Foley*, an appellate court ruled in February 2012 that the True Allele methodology was not novel (testimony at trial provided by Dr Mark Perlin) and further rejected the defendant's claim for the source code – 'scientists can validate the reliability of a computerized process even if the "source code" underlying that process is not available to the public'.

5. Defence attorneys invoked the Confrontation Clause in support of their argument that the FST source code is necessary to confront witnesses at trial. However, in New York State it is firmly established that evidence representing DNA testing procedures and results is non-testimonial in nature. See *People v Brown*, 2009 NY Slip OP 8475 (NY 2009), *People v Freycinet* 2008 NY Slip OP 5776 (NY 2008), *People v Rawlins*, 2008 NY Slip Op 1420 (NY 2008).

* These errors were reported in the document 'Release notes for version 2.2.1' in the notes in the section titled 'Version 1.2.4 released May 18, 2014' at http://scieg.org/lab_retriever.html.

† We gratefully acknowledge that much of this was provided by Melissa Mourges.

We suspect that a request for the code is not intended to obtain the code but rather to get a refusal, which can in itself be used as evidence. STRmix is looking at ways to disclose the code and still mitigate the commercial risk.

Suppliers of commercial code are reticent to risk disclosure. What realistic testing could someone do with the code other than compile it and run it? However there is a real risk of commercial damage. It is clear that defence should have meaningful access to a method for checking software. Would not an executable version be better? In fact an executable version that outputs intermediate values and access to the formulae is much more useful, in our opinion, than the code. We note the risk of variants of OSS proliferating and fragmenting the community. We already have Lab Retriever and LikeLTD as differing variants evolving separately from the same origin and introducing errors. Chris Steele makes the perfectly valid point that fragmentation has actually occurred without any assistance from OSS. There are now quite a number of independently created software packages. He goes on to argue, rationally, that this might lead to useful natural selection.

Ranked Lists of Weights: A Courtroom Discussion

Genotype	Weight
9,10	38.30%
8,9	24.66%
9,9	13.94%
7,9	6.67%
8,10	5.36%
10,10	5.23%
8,8	2.35%
7,10	1.29%
7,8	1.01%
Q,9	0.67%
Q,10	0.19%
7,7	0.13%
Q,8	0.11%
Q,7	0.10%

STRmix can give a ranked list of the weights for different genotype combinations. A hypothetical one is given above. We discuss here an argument we have met with in Australia: 'The POI is a 10,10 and he is not even in the top five possibilities'. We assume that this comment is made to suggest that the weight of evidence should be downgraded. The weight for 10,10 in this list is 0.0523. This will appear in the numerator of the *LR* for this locus. The smaller it is, the lower the *LR* for this locus. The evidential weight is being downgraded for this locus by the use of this term in the numerator. There is no need for any further adjustment. We assume that the comment about the placement in the list either arises from misunderstanding or the wish to create misunderstanding.

As a technical point the weight is the probability of the evidence given the genotype, not the probability of the genotype given the evidence. It is incorrect to call the 9,10 in this list the most probable genotype. If a description is needed, the best one is that the 9,10 is the genotype with the highest likelihood.

1.3 – clarifications

Point 1:

In the formula provided in the book chapter (chapter page 278) that reads:

$$LR = \frac{p(G_C \mid H_p, G_S, I)}{p(G_C \mid H_d, G_S, I)}$$

G_C refers to the crime scene profile (Genotype of Crime-scene). Gc is therefore a set of observed peaks with sizes and heights that are treated as random variables. It is true that the we may consider Pr(Gc), and typically the binary or semi-continuous methods will assign a probability. However, continuous systems may (and often do) exploit the fact that the parameters in both the numerator and denominator of the LR can be a ratio of densities, rather than strictly probabilities. Therefore, the initial term I the LR equation on chapter page 278 is specified as a density. In later LR equations on page 305 to 307, the similar terms used in the LR equation are given as probabilities. This is because if there is a difference in the number of contributors then the priors between the two terms are different and probabilities must again be used rather than densities.

Point 2:

On page 286 and 287 three are a number of integral formula (for example):

$$\int_M \Pr(\mathbf{O} \mid \mathbf{M}, \mathbf{S}_i) \Pr(\mathbf{M}) d\mathbf{M}$$

which should be formally written without the subscript under the integral term, for example correcting the above formula:

$$\int p(\mathbf{O}, \mathbf{M} \mid \mathbf{S}_i) d\mathbf{M}$$

Point 3:

The terms $\Pr(N = n \mid H_x)$ specify the prior belief by the party (prosecution or defence) that the number of contributors, *N*, takes any specific value, *n*, prior to seeing the profile. Typical practise in forensic genetics is to assign a probability of 1 to the *N* taking one value of *n* (for both Hp and Hd so that $\Pr(N = n \mid H_p) = \Pr(N = n \mid H_d)$) and 0 to all others. This removes the need for the summation across number of contributors altogether and the *LR* given in 9.1 simplifies to:

$$LR = \frac{\Pr(E \mid N = n, Hp)}{\Pr(E \mid N = n, Hd)}$$

And the explicit reference to number of contributors in the formula is dropped as the propositions are expected to possess this information (i.e. they are in the form, '*The DNA has originated from the POI and 2 other people*' hence requiring that *N* = 3), so that the standard LR formula is.

$$LR = \frac{\Pr(E \mid Hp)}{\Pr(E \mid Hd)}$$

In this section of the text we now consider a situation where the propositions need not specify a specific number of contributors, i.e. they can take the form '*The POI is (not) a contributor of DNA to the sample*'. We can then consider two possible treatments of the problem, either a single value for *N* is chosen, but can be different between the two parties so that $\Pr(N = n \mid H_p) \perp \Pr(N = n \mid H_d)$. We could them write the LR as:

$$LR = \frac{\Pr(E \mid N = n, Hp)}{\Pr(E \mid N = m, Hd)} \qquad n \in \mathbb{Z}^+, m \in \mathbb{Z}^+$$

Where *n* and *m* are used to specify that the two numbers can (but need not necessarily) be different. An alternative is to consider a range of values for *N*. Again, there is no need for the range being considered to be the same for the two parties i.e. prosecution may state the DNA has originated from 1 to 2 people and the defence may specify it has originated from 1 to 3. I agree with the examiner that in a formal sense there is no difference between equations 9.2 and 9.3. The change in summation indices was more to visually explain to a non-mathematical audience the assumptions being made within the formula itself. A formula more formally expressed would have been:

$$LR = \frac{\sum_n \Pr(E \mid N = n, Hp) \Pr(N = n \mid Hp)}{\sum_n \Pr(E \mid N = n, Hd) \Pr(N = n \mid Hd)} \qquad n \in \mathbb{Z}^+$$

Where the difference in ranges would be handled by the $\Pr(N = n \mid H_x)$ terms, e.g. if a uniform prior was used for $\Pr(N = n \mid H_x)$ then the example described above would be handled by considering:

$$\Pr(N = n \mid Hp) = \begin{cases} 1/2 & n \in \{1, 2\} \\ 0 & otherwise \end{cases}$$

$$\Pr(N = n \mid Hd) = \begin{cases} 1/3 & n \in \{1, 2, 3\} \\ 0 & otherwise \end{cases}$$

Due to software limitations, when a range can be specified for *N*, it is usual that the same range must be used for both parties.

There may be some instances where the case scenario being put forward could inform this probability, i.e. For the DNA profile produced from an intimate swab taken as part of a sexual assault where the victim has had no recent previous sexual contact with anyone, there may be more probability placed on *N* = 1or 2 (i.e. the victim, or victim and suspect, respectively). In the case of a swab of drug paraphernalia taken from the scene of a share house, there may be more probability placed on higher values for *N*. However, given difficulties in translating these situations into numerical values, it is often the case that equal probabilities are used for all values of *N* within the range **n** (where **n** is the set of contributor numbers being considered):

$$\Pr(N = n \mid Hp) = \Pr(N = n \mid Hd) = \frac{1}{|\mathbf{n}|} \qquad \text{for all } n$$

So that the *LR* is given by:

$$LR = \frac{\sum_{n \in \mathbf{n}} \Pr(E \mid N = n, Hp)}{\sum_{n \in \mathbf{n}} \Pr(E \mid N = n, Hd)} \qquad n \in \mathbb{Z}^{+}$$

As per eq (9.3).

Chapter 2: Models in the fully continuous interpretation system

The first task when creating a system that can be used to analyse DNA profiles is to describe how DNA profiles look and behave in the language of mathematics. During the development of STRmix™, before any programming, the most important features of DNA profiles, and how they could be described in real world terms, were determined i.e. the amount of template DNA, the level of DNA degradation and the efficiency with which the DNA profiling process occurred. There is little ability to short-cut this modelling the process, i.e. if peak height information is to be used in the analysis then it is a necessity to mathematically describe enough behavioural properties of DNA that the majority of peak fluorescence can be described. The assumption that is then made, is that any difference between the peak heights that are expected (from the models) and those that are observed (in the profile) is due to some system of stochastic peak height behaviour, which is then modelled using a peak height variability model.

The sub-sections in this chapter comprise the publications that explain different models used to describe DNA profile behaviour. The final publication in the chapter brings the models together into an MCMC based system and applies them to forensic problems. People most directly associate this last publication with the software STRmix™ as it is the heart of how the program works.

The various models used in STRmix™, or any DNA interpretation system, are usually grouped into categories; biological models and statistical models and it is worth briefly explaining how the two differ. The starting premise is that events occur in nature that are unable to be directly observed. Inferences are made that they have occurred because the effect of the event is seen in the DNA profiles that are produced. For example, it can be seen that peak heights tend to decrease across a DNA profile as molecular weight increases and so it is stated that degradation has been 'seen'. Of course, no degradation has been seen directly occurring (i.e. an analyst has not looked down a microscope and seen strands of DNA breaking apart before their eyes, or even cast their gaze over the wreckage of some DNA fragments, surmising it must have once been whole) but rather what is seen is the manifestation of the degradation in the particular type of data that has been generated. The mathematical process used to link the real-world event of degradation to the way it manifests itself on the EPG is considered a biological model. There is therefore a biological model that exists for every aspect of an EPG being described. Having biological models for various real-world events, there may be a desire to ask questions of the data, '*How much DNA is there from each contributor?*', '*Could Mr X be a contributor to this profile?*', '*Is this profile from 2 or 3 contributors?*', or '*Is this small peak an allele or an artefact?*'. These questions are not describing a biological event, but rather seeking some information. In order to translate the DNA profile data that has been obtained into answers for these questions requires the use of statistical models. For example, a statistical LR model is used to address questions of support for nominated individuals being donors to a DNA mixture. The statistical model of MCMC is used to glimpse at the posterior distribution of various parameters of interest, such as the amount of DNA each contributor has provided, or the goodness of fit of various genotypes in describing the peaks observed. As is often the case in science there are models within models and the picture can become complex as one seeks finer

and finer resolution, however dichotomising models into two types assists in the understanding of the whole picture.

Some of the papers in this chapter repeat material from the book chapter in the first chapter. The information has been supplied in the reverse order to how it was created. First came the more comprehensive mathematical descriptions, given in the papers in this chapter, and afterwards a gentler version produced for publication in DNA Evidence Interpretation.

2.1: The need to develop models to describe DNA profile behaviour

Start with the obvious trend that as more DNA is added to a system, then the resulting height of peaks on an EPG will be greater. Also, that when individuals possess the same alleles that a combination of their DNA result in a relative summation of their individual peaks to a single, indecomposable, larger version in the mixture. These concepts were already so well recognised prior to the development of STRmix™ that there was no ability to publish work in the area.

Stutter, too, was well recognised. The standard method for obtaining the expected height of the stutter was to regress the allelic designation (which is based on its molecular weight, or size) against stutter ratio (using data from a large validation study) and then having obtained the expected stutter ratio for the allele of interest, multiplying it by the parent peak height to obtain the expected stutter peak height. This model was found to work well for some loci and mediocrely for others. A simple linear regression model was implemented into the early version of STRmix™, noting that a better system did exist. That better system is the subject of the publication in section 2.2, which used the underlying sequence of the allele rather that its absolute size. This was called the 'LUS' model, for Longest Uninterrupted Sequence. The model was later refined even further to the 'multi-LUS' model, which is described in the paper in section 2.5. The multi-LUS model has since been incorporated into STRmix™ and validation work found that the statistic used to gauge how well the observed data is being described, improved markedly with the change. While reporting the refinement of the stutter model in the publication in section 2.5, we also took the opportunity to publish examples of how different aspects of DNA profile behaviour could be validates. We had, by this point, carried out these tasks a number of times when assisting laboratories with their validations, and had developed a number of simple and standardised methods.

Another important feature of DNA profile behaviour is the manner in which DNA degrades. In the most basic of descriptions, peaks get less intense as DNA fragment size increases. Imagine DNA as a long, wet noodle, the longer the strand the more prone to breakage. The simplest model to describe such a downward trend is linear, and this is indeed how degradation was initially modelled. The linear description of degradation worked well for the DNA profiling kits available at the time STRmix™ was first introduced. These were relatively simple (by current standards) profiling kits that had a size range of the DNA fragments produced from approximately 100 base pairs to 300 base pairs. In Australia and New Zealand in 2012 and 2013 (just after STRmix™ was introduced into active casework) there was an agreement by the Biology Specialist Advisory Group (a group consisting of senior member of government forensic biology laboratories around Australia and New Zealand) to increase the core number of loci from 9 to 18. This meant the introduction of new DNA profiling systems that possessed approximately double the number of regions (loci) examined. The DNA profile 'real estate' in the simple profiling systems was already very highly utilised, and so the introduction of more loci could only occur by two means, the addition of loci on a new dye channel or the extension of the current dye channels out past 300bp. The new kits did both. The consequence of the larger fragments was that the linear model of degradation was being extrapolated out to areas that hadn't previously been examined, and it was found lacking in some instances. Investigation into degradation models found that as the molecular weight range increased, a better

description of the degradation was achieved using an exponential distribution, something that is shown in the publication in section 2.3. STRmix™ was updated shortly after the publication to utilise the exponential degradation model.

These models of DNA profile behaviour, when combined, give the ability to describe what we would expect an electropherogram to look like, if we knew the values of various real-world parameters (such as DNA amount or level of degradation) values were. Of course, for evidence samples these parameter values are never know, but having models to translate them from parameter values to expected profiles allows a system to be developed that can be used to interpret DNA profile data, and this is described in the paper in section 2.6.

2.2: Stutter

Manuscript: Developing allelic and stutter peak height models for a continuous method of DNA interpretation. JA Bright, D Taylor, JM Curran, JS Buckleton. (2013) Forensic Science International: Genetics 7 (2), 296-304 – *Cited 51 times*

Statement of novelty: The idea of using the longest uninterrupted repeat had been previously published. This paper extends those works by carrying out a much more in-depth modelling of stutter ratio using the LUS system and provides an assessment of the model performance.

My contribution: I was a co-contributor to this work in the modelling and writing of the paper.

Research Design / Data Collection / Writing and Editing = 25% / 10% / 25%

Additional comments:

Forensic Science International: Genetics 7 (2013) 296–304

Contents lists available at SciVerse ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

Developing allelic and stutter peak height models for a continuous method of DNA interpretation

Jo-Anne Bright [a,b,*], Duncan Taylor [c], James M. Curran [b], John S. Buckleton [a]

[a] ESR Ltd, Private Bag 92021, Auckland, New Zealand
[b] Department of Statistics, University of Auckland, Private Bag 92019, Auckland, New Zealand
[c] Forensic Science South Australia, 21 Divett Place, SA 5000, Australia

ARTICLE INFO

Article history:
Received 12 May 2012
Received in revised form 3 November 2012
Accepted 29 November 2012

Keywords:
DNA interpretation
Mixture interpretation
Continuous models
Stutter

ABSTRACT

Traditional forensic DNA interpretation methods are restricted as they are unable to deal completely with complex low level or mixed DNA profiles. This type of data has become more prevalent as DNA typing technologies become more sensitive. In addition they do not make full use of the information available in peak heights. Existing methods of interpretation are often described as binary which describes the fact that the probability of the evidence is assigned as 0 or 1 (hence binary) (see for example [1] at 7.3.3). These methods are being replaced by more advanced interpretation methods such as continuous models. In this paper we describe a series of models that can be used to calculate expected values for allele and stutter peak heights, and their ratio *SR*. This model could inform methods which implement a continuous method for the interpretation of DNA profiling data.

1. Introduction

The forensic examination of biological evidence often produces low level or mixed DNA profiles, which are regarded as complex profiles. Traditional methods of interpretation are often described as binary which describes the fact that the probability of the evidence is assigned as 0 or 1 (hence binary) (see for example [1] at 7.3.3). These methods are being replaced by more advanced interpretation methods such as continuous models [2,3]. In this paper we describe a series of models that can be used to calculate expected values for allele and stutter peak heights, and their ratio, *SR* [2,3]. This is motivated by the difficulties traditional methods have with the interpretation of complex profiles [4,5]. Complicating interpretation of any DNA profile is the occurrence of stutter, an artefact of the PCR amplification of STR loci.

The earliest forms of the binary model considered alleles to be present or absent. Methods were subsequently developed that used heterozygous balance (*Hb*) to determine whether combinations of genotypes were supported or not. The binary model assigns a value of zero or one to the probability of the profile given the proposed allelic combination (hence the term binary) depending on whether the alleles could pair given *Hb*. The application of this model makes a number of assumptions including that peak area/height (hereafter height) is proportional to the quantity of template DNA and that the height of 'shared' peaks between individuals is the sum of the peaks from the contributing individuals. This is actually a rephrasing of the assumption that the height of a peak is linearly related to the quantity of DNA. Known shortcomings of the binary model [6,7] have led to the development of new and improved models that factor in the probability of drop-out [8–11]. Subsequently, fully continuous interpretation models have been developed [3,12]. These models take the quantitative information from the electropherogram (for example peak heights) and use them to calculate the probability of the peak heights given all the possible genotype combinations for the individual contributors. This approach removes some of the criticism regarding subjectivity [13,14] in profile analysis and attempt to ensure consistency in DNA interpretation and reporting across different laboratories. Well described probabilistic systems give a detailed accounting of their respective methods. What transpires inside a human expert's mind can be far more opaque than equations provided in peer-reviewed journals.

Continuous methods make assumptions about the underlying behaviour of peak height, or of the variability in the ratio of the two peaks of a heterozygote (*Hb*), and the ratio of allelic peak height to stutter peak height (*SR*) to evaluate the probability of a set of peak heights. These models may be developed from empirical data external to the profile under interpretation, by a combination of external data and the profile under consideration, or simply by the profile under consideration. We would tend to favour the combination approach.

In this paper we investigate the underlying behaviour of *Hb* and *SR*. We also investigate the relationship between the heights of two alleles of a heterozygote, and the allele and its stutter product. The

* Corresponding author at: ESR Ltd, Private Bag 92021, Auckland 1142, New Zealand. Tel.: +64 98153940; fax: +64 98496046.
E-mail address: Jo.bright@esr.cri.nz (J.-A. Bright).

1872-4973/$ – see front matter © 2013 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.fsigen.2012.11.013

aim is to build models to inform a continuous interpretation system. Previous work has investigated the variability in *Hb* in Applied Biosystems' Identifiler™ [15] and MiniFiler™ [16] multiplexes. The continuous model may work by means of modelling the variability in *Hb* directly but more often works with variability in peak heights themselves [2,3]. In single source profiles, the variability in *Hb* reduces as the average peak height (*APH*) at a locus increases.

The distribution of peak heights varies with the quantity of DNA and is difficult to investigate directly. The investigation could be undertaken by making consistent extractions and amplifications of entirely equivalent templates. In this case we would expect the height of each peak to vary about the same mean. The distribution could be determined directly. However the consistent replication of extraction and amplification template presents some experimental challenges. We are incapable of standardising the template to absolute precision. It is likely that the replicate peak heights would vary about a mean that was also varying. This is because template would vary and then the PCR process would add further variance. Since the two alleles of a heterozygote are as close as we can envisage to replicate extractions and amplifications of the same template, the variation in (the logarithm of) *Hb* should be twice that of (the logarithm of) peak height. This suggests that one practical route into modelling the distribution of peak height is through the distribution in *Hb*.

The variability in *SR* is routinely estimated by individual laboratories as part of an internal validation of a new multiplex or an analysis platform. Previous work has investigated the longest uninterrupted sequence (*LUS*) as a predictor of stutter [17,18]. It has been shown that alleles with large *LUS* values stutter more than alleles with small *LUS* values and plausibly amplify less. For any given *LUS* there will still be stutter peaks above or below expectation. A larger than expected stutter is likely to be caused by stutter events early in the PCR process. This would be expected to lead to a smaller allelic peak. This allows us to define the following hypothesis: If, for any given allele, the stutter peak is above expectation given its *LUS* value, then we expect the peak height for that allele to be below expectation. If this hypothesis were true, then this would have implications for any continuous model that sought to model stutter as well as allelic peaks independently.

Many laboratories are moving to the European Standard Set of Loci (ESSoL). One of the multiplexes which include these loci is Applied Biosystems' NGM SElect™. We report here an investigation into the variability of *Hb* and *SR* in this multiplex. We acknowledge that the concepts are universal across many different STR multiplexes. We have developed a biological model that can easily be grasped by a forensic biologist that is intended for use within any software implementing a continuous interpretation method.

2. Method

289 single source DNA profiles were analysed using Applied Biosystems' NGM SElect™ (Life Technologies, Carlsbad, CA) multiplex. The samples were saliva stains on FTA® Elute card (Whatman, Maidstone, England) and DNA was recovered off the card using a simple elute method. Prior to amplification all samples were quantified using Applied Biosystems' Quantifiler™ (Life Technologies, Carlsbad, CA) according to the manufacturer's instructions. A target of 1 ng of DNA was amplified using NGM SElect™ following the manufacturer's instructions in a 9700 silver block thermal cycler. Amplified products were separated on an Applied Biosystems' 3130xl Genetic Analyser (Life Technologies, Carlsbad, CA) and data was analysed using Applied Biosystems' GeneMapper™ *ID* version 3.2.1 (Life Technologies, Carlsbad, CA) using a 25 RFU limit of detection threshold.

Loci where the alleles were separated by one repeat were discarded because stutter is likely to interfere with the allele height of the low molecular weight allele in an additive manner. These have previously been referred to as *stutter affected heterozygotes*. In total, 2323 heterozygous loci were identified as being suitable for analysis.

Stutter ratio was defined as

$$SR = \frac{O_{a-1}}{O_a}$$

where O_{a-1} refers to the observed height of the stutter peak, and O_a the parent peak.

LUS was defined as the longest stretch of basic repeat motifs within the allele. The longest uninterrupted sequence (*LUS*) for each allele was determined using the method of Brookes et al. [17]. *LUS* values were obtained by looking up the allele designation in the short tandem repeat DNA internet database (STRBase) [19,20]. Where multiple values for *LUS* were available the average *LUS* value across the reported variants observed was taken. δ_{LUS} was defined as the difference in *LUS* values for the two alleles of a heterozygote. Heterozygote balance (*Hb*) was calculated as

$$Hb = \frac{O_H}{O_L}$$

where O_H refers to the height of the high molecular weight allele and O_L the height of the low molecular weight allele. Statistical analysis was undertaken using R [21] and MS EXCEL™.

Linear modelling was used to test the effect of various explanatory variables on the expected values of *SR* and *Hb*. Having chosen a model for the expected value we investigate models to predict the variance about this expectation.

3. Results

3.1. Stutter

The following linear model was proposed to describe the relationship between *SR* and the explanatory variables *LUS* and locus, *l*:

$$SR_i = \beta_{0,l} + \beta_{1,l} LUS_i \quad (1)$$

This was termed the stutter model. Linear modelling of stutter has been reported previously [22,23]. The model described in this paper was selected after exploratory analysis suggested a nil or small effect of other potential explanatory variables. The plots of *SR* versus *LUS* for individual NGM SElect™ loci are given in Appendix 1. The interaction term allows a different slope of the *SR* versus *LUS* line for each locus. The R^2 is 0.83 for the stutter model. The improvement in fit of *LUS* over simple allele number is demonstrated for the TH01 locus in Figs. 1 and 2 where Fig. 1 gives *SR* versus repeat number ($R^2 = 0.02$) and Fig. 2 *SR* versus *LUS* ($R^2 = 0.58$). Of interest may be the 9.3 allele in TH01. This has the structure $[AATG]_6ATG[AATG]_3$ and hence has a LUS of 6. Inspection of Figs. 1 and 2 show that the 9.3 allele sits much better in the trend when placed at an *LUS* of 6.

We would anticipate that log(*SR*) would be easier to model because *SR* is a ratio. Given that the allelic peak height is much bigger than the stutter peak height this effect should be minor.

There is good support for using a linear relationship to model the behaviour of *SR* with respect to *LUS* (see Appendix 1). In addition, *SR* is a standard concept for forensic biologists and so avoiding the introduction of logarithmic scales will improve model acceptance. A summary of the intercepts and slopes, using this model, for every locus in the NGM SElect™ multiplex kit is given in Appendix 2. D2S441 is very poorly described by this model.

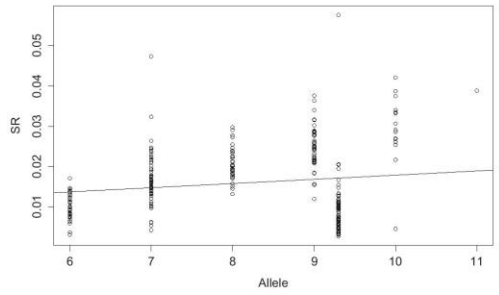


Fig. 1. A plot of stutter ratio versus allele repeat number for the TH01 locus.

A normal quantile–quantile (Q–Q) plot of the residuals from the model versus theoretical quantiles from a normal distribution is presented in Fig. 3.

The Q–Q plot suggests that the data is symmetric but with heavier tails than the normal distribution. An assumption of approximate normality is plausibly acceptable noting that there are a great many data points in the central region.

The squared residuals were regressed against allele height in order to investigate the factors affecting the variability of *SR*. There is a significant effect of allele height on the variance of *SR* ($p = 3.9 \times 10^{-14}$) however, as the coefficient was small (-7.5×10^{-7}), it will have little effect on the predicted variability of *SR*.

Some alleles show markedly larger variation in *SR* compared with the expectation. For example, at locus D2S441 the *SR* for several values of *LUS* are not well described by the model (refer Appendix 1). Closer inspection suggests that, in many cases, this was caused by an allele that has a complex repeat structure comprising of variant regions with differing *LUS* values. In another example, for D21S11 30 the sequence has been variously typed as:

$[TCTA]_6\ [TCTG]_5\ [TCTA]_3\ TA\ [TCTA]_3\ TCA\ [TCTA]_2\ TCCA\ TA\ [TCTA]_{11}$
$[TCTA]_5\ [TCTG]_6\ [TCTA]_3\ TA\ [TCTA]_3\ TCA\ [TCTA]_2\ TCCA\ TA\ [TCTA]_{11}$
$[TCTA]_4\ [TCTG]_6\ [TCTA]_3\ TA\ [TCTA]_3\ TCA\ [TCTA]_2\ TCCA\ TA\ [TCTA]_{12}$
$[TCTA]_6\ [TCTG]_6\ [TCTA]_3\ TA\ [TCTA]_3\ TCA\ [TCTA]_2\ TCCA\ TA\ [TCTA]_{10}$

for different variants [19]. Since we will only know the molecular weight and not the sequence when using typical casework electropherograms we have used an average *LUS*. In this case we have used the average of 10, 11, 11 and 12. Since we have used an average and the sample plausibly contains some of each sequence we expect to see enhanced spread.

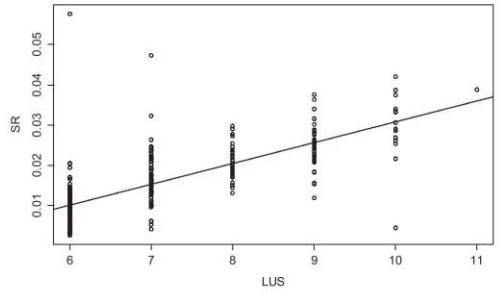


Fig. 2. A plot of stutter ratio versus *LUS* for the TH01 locus.

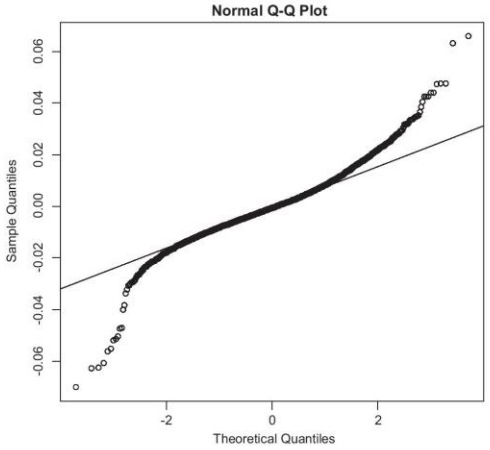


Fig. 3. A plot of the Q–Q plot from the full stutter model.

If there is indeed an effect of *LUS*, as observed here and previously [17,18], then, using the D21S11 example given above we would expect some variants in our set with LUS values of 10, 11 and 12. This would lead to distributions in the observed stutter ratio that are centred around a higher value (for *LUS* = 12) and a lower value (for *LUS* = 10) but all are plotted at *LUS* = 11. Hence a wider spread. Such widening is a likely explanation for the heavy tails observed in Fig. 3.

3.2. Heterozygote balance variability

The relationship between *Hb* and average peak height (*APH*) was demonstrated for NGM SElect™ data in Fig. 4. The variation in *Hb* decreases as *APH* increases. This funnel shape has been observed in other multiplexes [15,16]. Direct comparison of the distributions shows that there is less variation in *Hb* with NGM SElect™ compared with that seen in the Identifiler™ and MiniFiler™ multiplexes [15,24].

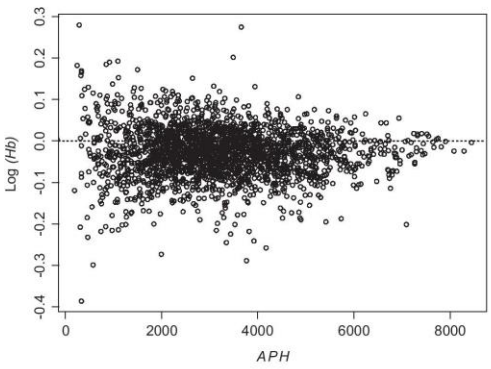


Fig. 4. log(*Hb*) versus *APH*, 2323 heterozygote NGM SElect™ loci.

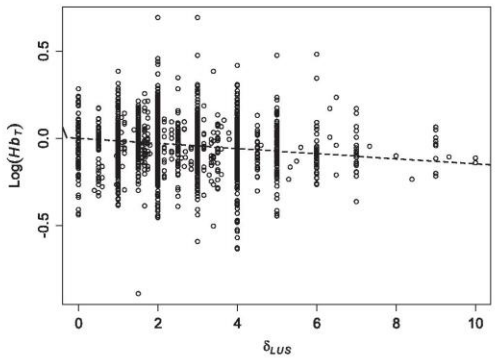


Fig. 5. A plot of $\log(Hb_T)$ versus δ_{LUS}.

It is known [17,18], and reinforced above, that alleles with large *LUS* values stutter more. One would expect alleles with large *LUS* values to have smaller allelic peaks [7] for a given template level. Under this hypothesis, stuttering is one of the determinants of any systematic effect on *Hb*, and it is the difference in *LUS*, δ_{LUS}, that should be the explanatory variable for *Hb*. Since *Hb* is a ratio we expect $\log(Hb)$ to be more amenable to modelling. This is supported by previous work [15,16]. It is helpful to consider the concept of the sum of the allelic and stutter peaks, termed total allelic product (T) [7]. This is calculated, for the ath allele, as

$$T_a = O_{a-1} + O_a \tag{2}$$

We can now define Hb_T in terms of total allelic product:

$$Hb_T = \frac{T_H}{T_L} \tag{3}$$

where T_H and T_L are the total allelic product values for the high and low molecular weight alleles respectively. If stutter is the only cause of variation in allelic peak height within the two peaks of a heterozygote, then we expect the mean of $\log(Hb_T)$ to be zero, and to have no relationship with δ_{LUS} or any other variable. In Fig. 5 we give the plot of $\log(Hb_T)$ versus δ_{LUS}. The regression line was forced through the origin. There was a small but significant negative slope to the regression line in Fig. 5 (slope = −0.0047). A plot of $\log(Hb_T)$ versus the difference in allele repeats (S_{AR}) also has a small but significant downwards slope (slope = −0.0053, data not shown). We conclude from this that there is something other than just stutter affecting allelic peak height for a given template level. This is likely to be simply due to the reduced amplification efficiency of the larger allele at a heterozygote locus. Of course template level is the primary determinant of peak height but should have no effect on expected Hb_T. After template the next largest effect appears to be stutter ratio and this affects both *Hb* and peak heights, but should not affect Hb_T. Last there is something else which we, and others, postulate is simply amplification efficiency. This affects peak heights, *Hb* and Hb_T. We are unable to determine from this analysis whether the behaviour of this last effect, postulated as relative amplification efficiency is better predicted by δ_{AR} or δ_{LUS} however both exhibit a small but significant effect on Hb_T.

3.3. Modelling peak heights

In this section we model peak heights as opposed to the ratios *Hb*, Hb_T and *SR*. In order to develop a model for expected peak height we need first to model the expected value for true mass at each allelic position at a locus T^l_{an}. We coin the term mass to subsume considerations of template, degradation and locus amplification effects. The 'true' mass of template DNA is not known. We model mass based on our observations of the data and understanding of the behaviour of DNA profiles. During modelling of peak heights versus molecular weight for various multiplexes we have observed that some are adequately explained with a linear model whereas some require an exponential model. NGM SElect™ appears to be adequately modelled using the simpler linear model.

For L loci, N contributors and R replicates the height of an allele, a, at locus l, for replicate r, from contributor n is modelled as:

$$T^l_{anr} = A^l_{.r}(t_n + d_n \times m^l_a)X^l_{an} \tag{4}$$

where m^l_a is the molecular weight of allele a at locus l; $A^l_{.r}$ ($l = 1 \ldots L$, $r = 1 \ldots R$) is the locus offset at locus l, replicate r; t_n ($n = 1 \ldots N$) is the intercept of the line for mass versus molecular weight for contributor n; d_n ($n = 1 \ldots N$) is the slope of the line for mass versus molecular weight for contributor n; X^l_{an} is the count of allele a at locus l in contributor n. $X^l_{an} = 1$ for a heterozygote with allele a and $X^l_{an} = 2$ for a homozygote a.

We refer to the variables A, t, and d collectively as the mass variables **M**. Note that when considering one amplification of a sample we can drop the 'r' subscripts, which we subsequently do so for simplicity. The locus offset, A^l, allows different amplification efficiencies for each locus. One A^l value may be set arbitrarily, termed 'fixed' and the others allowed to vary, termed 'free'. If A^l is allowed to be completely free it will tend to the midpoint of a heterozygote for single source profiles and to a related position for mixtures. This is unacceptable and would impose a large negative correlation between the peak height residuals. Accordingly we set the probability of each of the $L - 1$ free locus specific amplification efficiency parameters A^l for each of the $L - 1$ loci as $N(\mu_A, \sigma_A)$ where μ_A is the simple arithmetic average of the A^l values and σ_A is a preset hypervariable. This allows a limited freedom to the A^l variables but penalises any single value that departs significantly from the average. We set a uniform prior on μ_A.

3.4. Application of the model for mass and stutter

Mass at an allelic position at a locus can be apportioned to stutter and allele using the following equations where *SR* is determined from the model.

$$E^l_{(a-1)n} = \frac{SR^l_a(T^l_{an})}{1 + SR^l_a} \tag{5}$$

$$E^l_{an} = \frac{T^l_{an}}{1 + SR^l_a} \tag{6}$$

where $E^l_{(a-1)n}$ is the expected stutter peak height of the ath allele for the nth contributor at locus l; E^l_{an} is the expected allelic peak height of the ath allele for the nth contributor at locus l.

Mass was assigned for each allele for a subset of 100 samples from the NGM SElect™ dataset. The subset included both heterozygote and homozygote loci but all stutter affected heterozygotes were removed. Mass variables A^l, t_n and s_n were determined by a maximum likelihood method.

The stutter model (Eq. (1)) was used to calculate the expected stutter ratio for each allele. Weusten and Herbergs [25] suggest that the relative standard deviation on the numbers of chains should be inversely proportional to the square root of the expected number of DNA strands entering the amplification. This suggests that the 95% standard error intervals on stutter ratio should have the shape $k/\sqrt{T^l_{an}}$. Fig. 6 is a plot of the logarithm of the ratio of observed and expected heights are plotted against T^l_{an}.

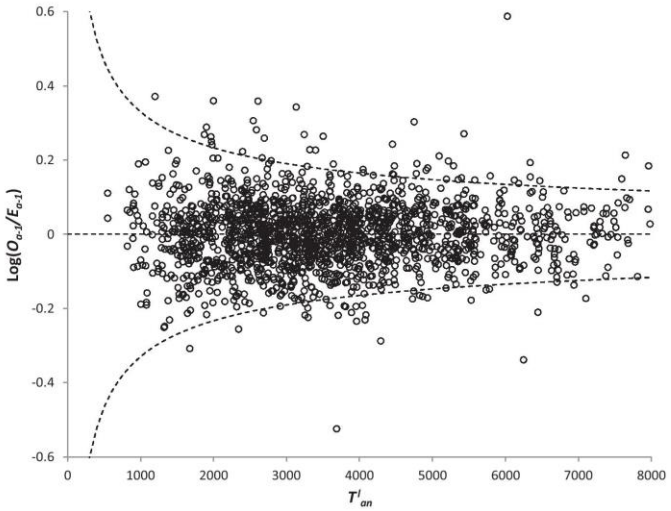


Fig. 6. A plot of log O_{a-1}/E_{a-1} versus T^l_{an} for the stutter peaks. The dotted lines approximate ±2 standard error intervals.

Subsequently the expected heights of the allele peaks were calculated for each sample. The variance of the allele model is examined in Fig. 7 where the logarithm of the ratio of observed and expected heights are plotted against T^l_{an}. The *x*-axis has been truncated in Fig. 7 at 8000 RFU to avoid saturation effects. At allele heights above approximately 8000 RFU, the data points tend to rise above the trend. These data points are likely to be affected by saturation of the 3130 camera, where the relationship between amount of DNA and allele height is no longer linear.

It has been suggested that the variance of the allele model (Fig. 7) is inversely proportional to the expected peak height, c_a^2/E^l_{an} [26]. The dotted lines are $\pm 1.96(c_a/\sqrt{E^l_{an}})$, where $c = 3.95$ fitted by MLE. These approximate ±2 standard error intervals are aimed at emphasising the shape of the model fitted to the data. Inspection of these plots indicates that the models are a reasonable description of the data, with few data points observed outside the intervals. The variance is symmetric around mean = 0.

Recall that expected height is developed from the mass variables, **M**. If the predicted T^l_{an} for each of the two alleles of a heterozygote using **M** is correct, then these two variables are conditionally independent given **M**. We could reasonably expect, then, that given **M** the log O_a/E_a value for each allelic peak of a

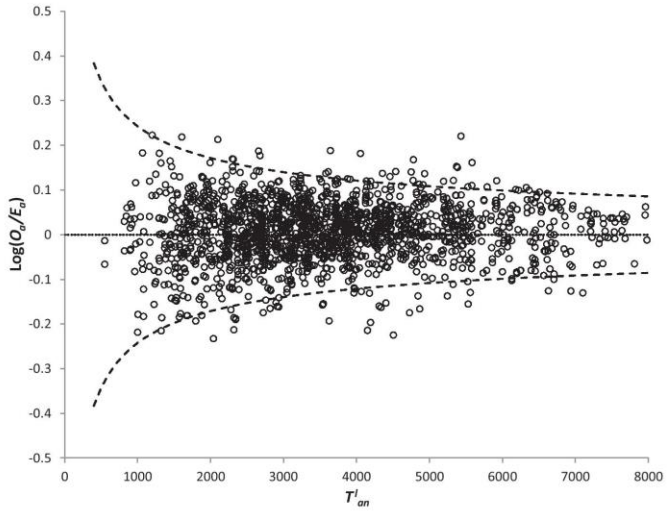


Fig. 7. A plot of log O_a/E_a versus T^l_{an} for the allelic peaks. The dotted lines approximate ±2 standard error intervals.

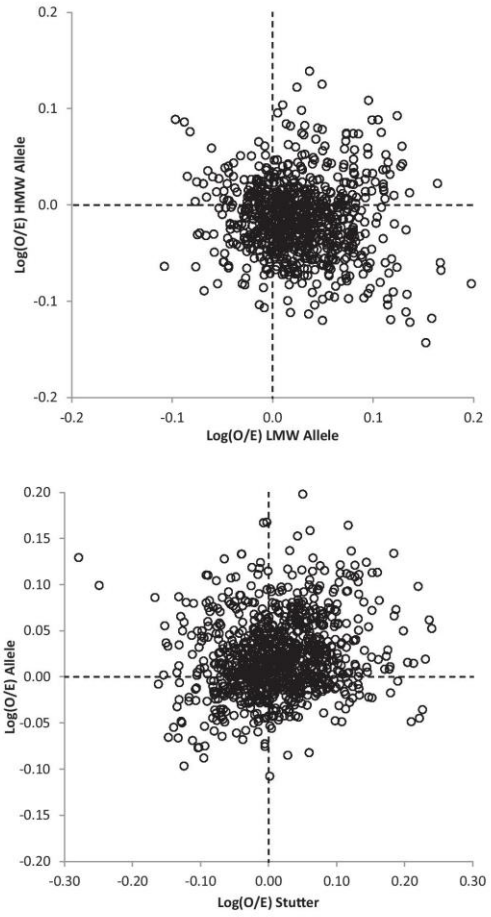


Fig. 8. (a) log O_H/E_H for the high molecular weight allele versus log O_L/E_L for the low molecular weight allele for each heterozygote locus and (b) log O_a/E_a for the allelic peak versus log O_{a-1}/E_{a-1} for stutter peak.

heterozygote is uncorrelated. However we would still anticipate a negative correlation between the log O_a/E_a values for the allele and log O_{a-1}/E_{a-1} for the associated stutter peak.

The correlation between the observed and expected peak heights at each heterozygote locus and between the observed and expected peak heights of allele and stutter was investigated, graphically (see Fig. 8a and b, respectively). The Pearson product-moment correlation coefficient was calculated as −0.0795 for log O_H/E_H for the HMW allele versus log O_L/E_L for the LMW allele and 0.1157 for log O_a/E_a allele versus log O_{a-1}/E_{a-1} stutter.

Unexpectedly the scatter plots in Fig. 8a and b indicate that there is no detectable correlation between stutter and allele in this biological model.

Assuming an approximate normal distribution, with a mean of zero, a constant variance for the stutter model, and variance = c_a^2/E_{an}^l for the allele model and variance = c_s^2/E_{an}^l for the stutter model then:

$$\frac{\log O_{(a-1)n}^l}{\log E_{(a-1)n}^l} \sim N(0, c_s^2/E_{an}^l)$$
$$\frac{\log O_{an}^l}{\log E_{an}^l} \sim N(0, c_a^2/E_{an}^l)$$

Plots to check for normality for the allele and stutter models indicate that the assumption of normality is sustainable (data not shown). Both tails of the distribution appear heavy. Additional exploratory modelling of the data (data not shown) including fitting a gamma distribution does not improve the fit.

4. Discussion

Previous publications have suggested that *LUS* is a better explanatory variable for *SR* than allele designation. This is confirmed for the NGM SElect™ multiplex. However one locus, D2S441 is very poorly described by this model. One plausible explanation is that the sequence data needs re-evaluation.

Weusten and Herbergs [25] have suggested that the 95% standard error intervals on stutter ratio should have the shape $k/\sqrt{T_{an}^l}$. This equation was plotted as dotted lines in Fig. 6, supporting the theory.

When considering the mean value of *Hb* we expect no effect of template although template is thought to affect the variance about this mean. Stutter ratio does have an effect on mean *Hb* especially when the alleles differ significantly in *LUS*. *SR* alone however is not the only factor in predicting mean *Hb*. This can be observed in Fig. 5. Larger alleles amplify less efficiently. This is likely to be due to an amplification effect with the longer lengths of DNA resulting in lower peak heights.

The concept of mass (T) was introduced in order to model allele heights and stutter heights. T was described by the molecular weight of the allele and the three mass parameters; amplification efficiency, intercept, and slope. In this research, mass parameters were determined using maximum likelihood. More elegant methods such as MCMC exist [27].

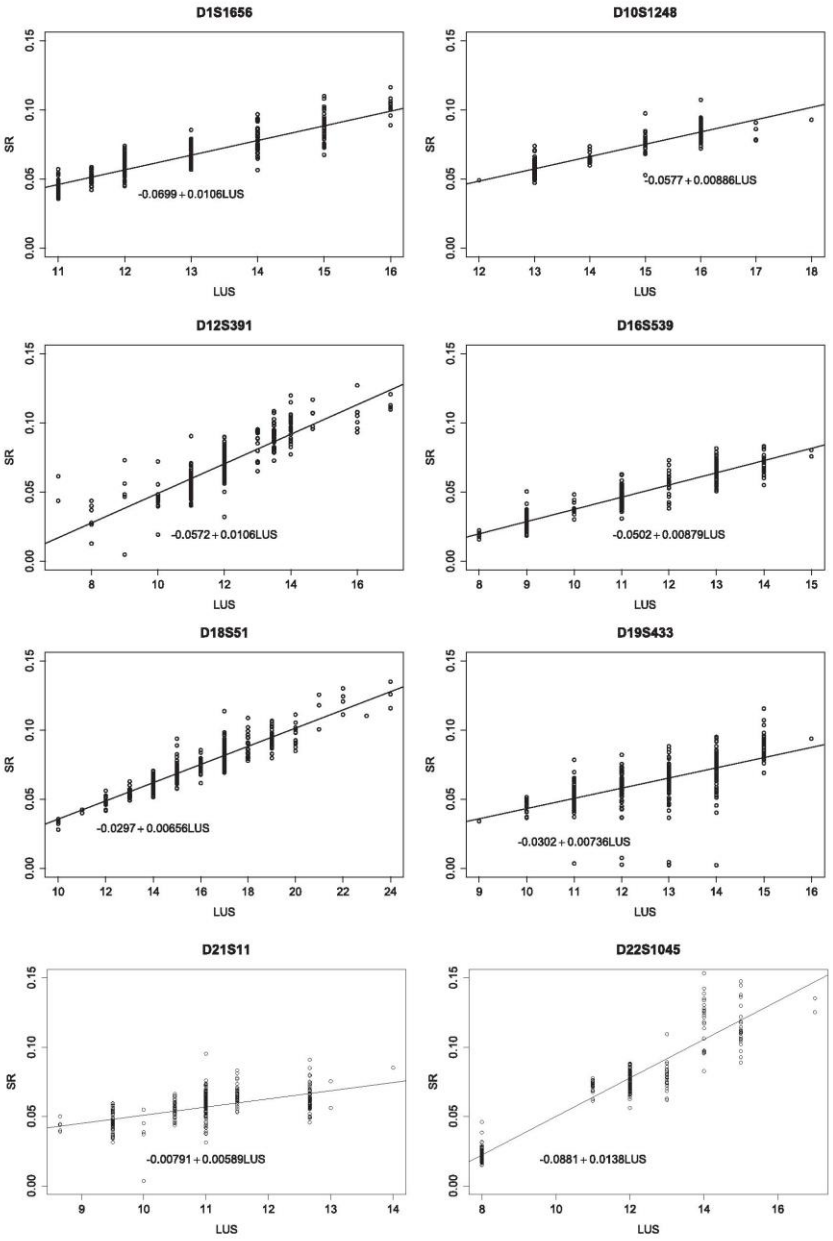
T_{an}^l and *SR* were combined to calculate expected heights for stutter and allele. The approximate linearity of the investigative plots showed an acceptable fit to the log normal distribution. Both tails appear heavy which does not suggest that the gamma models being considered by some commentators are a total solution [2,28,29]. The correlation graphs, Fig. 8a and b, show no detectable relationship between the expected heights of alleles and their corresponding stutter, and the *HMW* and *LMW* alleles at a heterozygous locus. This suggests that the independence model may be sustainable.

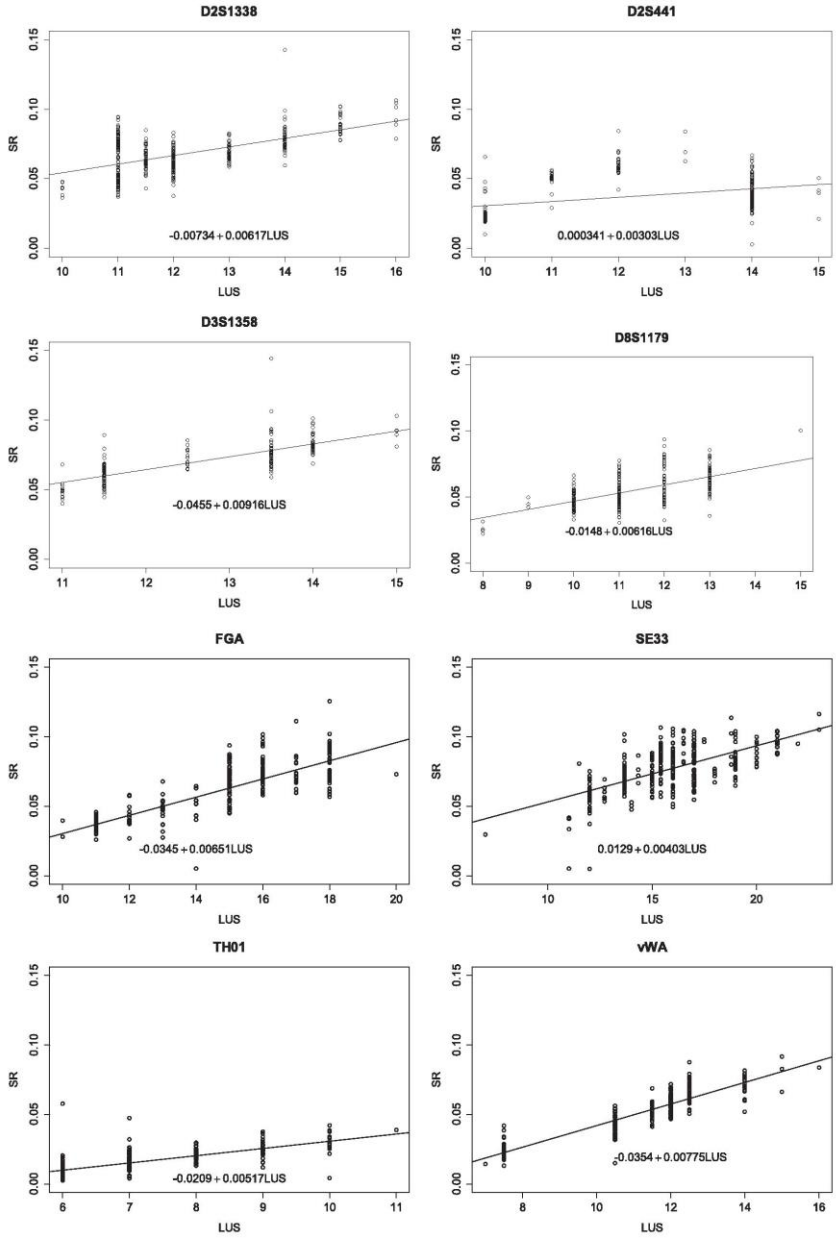
We have described a model that can be used to predict expected values and variances for *SR* but further give models for predicted allele and stutter heights and the variances about these predictions. We did not find a correlation between higher than expected allele peak and lower than expected stutter peaks.

Acknowledgements

This work was supported in part by grant 2011-DN-BX-K541 from the US National Institute of Justice. Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the U.S. Department of Justice. We gratefully acknowledge the comments of Catherine McGovern, Stuart Cooper and two anonymous reviewers which have greatly improved this paper.

Appendix 1. Stutter ratio versus *LUS* for individual NGM SElect™ loci.




D2S1338
-0.00734 + 0.00617LUS
D2S441
0.000341 + 0.00303LUS
D3S1358
-0.0455 + 0.00916LUS
D8S1179
-0.0148 + 0.00616LUS
FGA
-0.0345 + 0.00651LUS
SE33
0.0129 + 0.00403LUS
TH01
-0.0209 + 0.00517LUS
vWA
0.0354 + 0.00775LUS
SR
LUS

Appendix 2. Summary of the stutter model $SR_i = \beta_{0,l} + \beta_{1,l}LUS_i$.

Locus	Intercept	Slope
D10S1248	−0.0576	0.0089
D12S391	−0.0571	0.0107
D16S539	−0.0502	0.0088
D18S51	−0.0297	0.0066
D19S433	−0.0302	0.0074
D1S1656	−0.0699	0.0106
D21S11	−0.0079	0.0059
D22S1045	−0.0881	0.0139
D2S1338	−0.0073	0.0062
D2S441	0.0004	0.0031
D3S1358	−0.0455	0.0092
D8S1179	−0.0148	0.0062
FGA	−0.0344	0.0066
SE33	0.0129	0.0041
TH01	−0.0208	0.0052
vWA	−0.0354	0.0078

References

[1] T.M. Clayton, J.S. Buckleton, Mixtures, in: Forensic DNA Evidence Interpretation, CRC Press, Boca Raton, FL, 2004, pp. 217–274.
[2] R.G. Cowell, S.L. Lauritzen, J. Mortera, Probabilistic modelling for DNA mixture analysis, Forensic Sci. Int. Genet. Suppl. Ser. 1 (1) (2008) 640–642.
[3] M.W. Perlin, M.M. Legler, C.E. Spencer, J.L. Smith, W.P. Allan, J.L. Belrose, et al., Validating TrueAllele® DNA mixture interpretation, J. Forensic Sci. 56 (6) (2011) 1430–1447.
[4] J.-A. Bright, P. Gill, J. Buckleton, Composite profiles in DNA analysis, Forensic Sci. Int. Genet. 6 (3) (2012) 317–321.
[5] H. Kelly, J.-A. Bright, J. Curran, J. Buckleton, The interpretation of low level DNA mixtures, Forensic Sci. Int. Genet. 6 (2) (2012) 191–197.
[6] J. Buckleton, C.M. Triggs, Is the 2p rule always conservative? Forensic Sci. Int. 159 (2006) 206–209.
[7] J.S. Buckleton, C.M. Triggs, S.J. Walsh, Forensic DNA Evidence Interpretation, CRC Press, Boca Raton, FL, 2004.
[8] H. Haned, Forensim: an open-source initiative for the evaluation of statistical methods in forensic genetics, Forensic Sci. Int. Genet. 5 (4) (2011) 265–268.
[9] H. Haned, P. Gill, Analysis of complex DNA mixtures using the Forensim package, Forensic Sci. Int. Genet. Suppl. Ser. 3 (1) (2011) e79–e80.
[10] D.J. Balding, J. Buckleton, Interpreting low template DNA profiles, Forensic Sci. Int. Genet. 4 (1) (2009) 1–10.
[11] K.E. Lohmueller, N. Rudin, Calculating the weight of evidence in low-template forensic DNA casework, J. Forensic Sci. (2012).
[12] I.W. Evett, P.D. Gill, J.A. Lambert, Taking account of peak areas when interpreting mixed DNA profiles, J. Forensic Sci. 43 (1) (1998) 62–69.
[13] L. Geddes, What are the chances, New Sci. 207 (2774) (2010) 8–10.
[14] L. Geddes, C. King, C. Stier, Between prison and freedom, New Sci. 207 (2773) (2010) 8–11.
[15] J.-A. Bright, J. Turkington, J. Buckleton, Examination of the variability in mixed DNA profile parameters for the Identifiler™ multiplex, Forensic Sci. Int. Genet. 4 (2) (2009) 111–114.
[16] J.-A. Bright, E. Huizing, L. Melia, J. Buckleton, Determination of the variables affecting mixed MiniFiler™ DNA profiles, Forensic Sci. Int. Genet. 5 (5) (2011) 381–385.
[17] C. Brookes, J.-A. Bright, S. Harbison, J. Buckleton, Characterising stutter in forensic STR multiplexes, Forensic Sci. Int. Genet. 6 (1) (2012) 58–63.
[18] P.S. Walsh, N.J. Fildes, R. Reynolds, Sequence analysis and characterization of stutter products at the tetranucleotide repeat locus vWA, Nucleic Acids Res. 24 (1996) 2807–2812.
[19] J.M. Butler, D.J. Reeder, Short tandem repeat DNA internet database. Available from: www.cstl.nist.gov/biotech/strbase.
[20] C.M. Ruitberg, D.J. Reeder, J.M. Butler, STRBase: a short tandem repeat DNA database for the human identity testing community, Nucleic Acids Res. 29 (1) (2001) 320–322.
[21] R Development Core team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2004.
[22] M.W. Perlin, G. Lancia, S.-K. Ng, Toward fully automated genotyping: genotyping microsatellite markers by deconvolution, Am. J. Hum. Genet. 57 (1995) 1199–1210.
[23] S.C. Bakker, R.J. Sinke, P.L. Pearson, Differences in stutter intensities between microsatellites are related to length and sequence of the repeat, in: Unravelling the Genetics of Schizophrenia and ADHD, University of Utrecht, Utrecht, 2005, pp. 81–94.
[24] J.-A. Bright, E. Huizing, L. Melia, J. Buckleton, Determination of the variables affecting mixed MiniFiler(TM) DNA profiles, Forensic Sci. Int. Genet. 5 (5) (2011) 381–385.
[25] J. Weusten, J. Herbergs, A stochastic model of the processes in PCR based amplification of STR DNA in forensic applications, Forensic Sci. Int. Genet. 6 (2012) 17–25.
[26] H. Kelly, J.-A. Bright, J.M. Curran, J. Buckleton, Modelling heterozygote balance in forensic DNA profiles, Forensic Sci. Int. Genet. 6 (6) (2012) 729–734.
[27] J.M. Curran, A MCMC method for resolving two person mixtures, Sci. Justice 48 (4) (2008) 168–177.
[28] R.G. Cowell, S.L. Lauritzen, J. Mortera, Probabilistic expert systems for handling artifacts in complex DNA mixtures, Forensic Sci. Int. Genet. 5 (3) (2011) 202–209.
[29] R.G. Cowell, Validation of an STR peak area model, Forensic Sci. Int. Genet. 3 (3) (2009) 193–199.

2.3: Degradation

Manuscript: Degradation of forensic DNA profiles. JA Bright, D Taylor, JM Curran, JS Buckleton. (2013) Australian Journal of Forensic Sciences 45 (4), 445-449 – *Cited 23 times*

Statement of novelty: At the time of publication the work was the first to compare the performance of the different models of degradation (linear and exponential) acting on DNA profile data.

My contribution: I was a co-contributor to this work in the modelling and writing of the paper.

Research Design / Data Collection / Writing and Editing = 20% / 5% / 20%

Additional comments:

Australian Journal of Forensic Sciences

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tajf20

Degradation of forensic DNA profiles

Jo-Anne Bright [a b], Duncan Taylor [c], James M. Curran [b] & John S. Buckleton [a]

[a] ESR, Private Bag 92021, Auckland, 1025, New Zealand

[b] Department of Statistics, University of Auckland, Private Bag 92019, Auckland, 1025, New Zealand

[c] Forensic Science South Australia, 21 Divett Place, SA, 5000, Australia

Version of record first published: 14 Mar 2013.

To cite this article: Jo-Anne Bright , Duncan Taylor , James M. Curran & John S. Buckleton (2013): Degradation of forensic DNA profiles, Australian Journal of Forensic Sciences, DOI:10.1080/00450618.2013.772235

To link to this article: http://dx.doi.org/10.1080/00450618.2013.772235

Australian Journal of Forensic Sciences, 2013
http://dx.doi.org/10.1080/00450618.2013.772235

Degradation of forensic DNA profiles

Jo-Anne Bright[a,b]*, Duncan Taylor[c], James M. Curran[b] and John S. Buckleton[a]

[a]*ESR, Private Bag 92021, Auckland 1025, New Zealand;* [b]*Department of Statistics, University of Auckland, Private Bag 92019, Auckland 1025, New Zealand;* [c]*Forensic Science South Australia, 21 Divett Place, SA 5000, Australia*

Selected profiles typed at the Promega PowerPlex 21 (PP21) loci were examined to determine if a linear or exponential model best described the relationship between peak height and molecular weight. There were fewer large departures from observed and expected peak heights using the exponential model. The larger differences that were observed were exclusively at the high molecular weight loci. We conclude that the data supports the use of an exponential curve to model peak heights versus molecular weight in PP21 profiles. We believe this observation will improve our ability to model expected peak heights for use in DNA interpretation software.

Keywords: forensic DNA; PowerPlex 21; degradation

Introduction

In the interpretation of forensic DNA evidence a sample associated with a crime is compared with genotype information from one or more persons. Typically, the samples will be amplified using commercially manufactured short tandem repeat (STR) multiplexes that analyse many loci simultaneously, with subsequent polymerase chain reaction (PCR) product generated on a capillary electrophoresis instrument. The resulting DNA profile is an electropherogram (epg). The heights (or areas) of the peaks within the epg are approximately proportional to the amount of undegraded template DNA[1–4]. However this relationship is affected by a number of systematic factors. Notable amongst these factors is the molecular weight (m_a) of allele, a.

A typical epg has a downward trend with increasing molecular weight. This is variously described as the degradation slope or the 'ski slope'[5–7]. The term degradation slope alludes to a suggested cause, degradation of the DNA. There are many chemical, physical and biological insults that are believed to contribute to DNA degradation or inhibition of a profile. Environmental factors such as humidity[8], bacteria[7] or other forces such as ultraviolet light break down the DNA, destroying some fraction of the initial template[9]. Although the cause of the slope may not be known, we will refer to this ski slope effect as degradation to comport with common usage.

The modelling of expected peak heights is important in the interpretation of forensic mixtures. The authors have previously described a series of models that can be used to calculate expected values for allele and stutter peak heights, and their ratio, SR[10].

*Corresponding author. Email: jo.bright@esr.cri.nz

Known shortcomings of the binary model[11,12] have led to the development of new and improved models that factor in the probability of dropout[13–16]. Subsequently, fully continuous interpretation models have been developed[17,18]. These models take the quantitative information from the electropherogram (for example peak heights) and use them to calculate the probability of the peak heights given all the possible genotype combinations for the individual contributors. This approach removes some of the criticism regarding subjectivity[19,20] in profile analysis and attempts to ensure consistency in DNA interpretation and reporting across different laboratories.

It is important to understand how degradation affects these models. The simplest model is linear. That is, the expected peak height declines constantly with respect to molecular weight. This can be demonstrated crudely by taking a paper epg and drawing a downward sloping straight line across the apex of the heterozygote peaks from the lowest molecular weight locus to the highest molecular weight locus. A linear model has previously been suggested by the current authors[10]. Tvedebrink et al.[21] have proposed an exponential relationship when considering models for allelic dropout.

If the breakdown of the DNA strand was random with respect to location, then we would expect that the observed height of peak *a*, O_a, would be exponentially related to molecular weight. In this work we investigate linear and exponential equations for modelling degradation within single source Promega PowerPlex 21 profiles.

Methods

We analysed data from all Australian state and territory laboratories generated using the Promega PowerPlex 21 multiplex as part of a large data analysis project to implement a continuous model of DNA interpretation in Australasia.

Single source PowerPlex 21 (Promega Corporation, Madison, WI) DNA profiles were submitted for analysis from eight laboratories, either as previously analysed outputs or as raw, unanalysed data files. All raw data were analysed using Applied Biosystems' GeneMapper ID v 3.2.1 with an analysis threshold of 30 relative fluorescent units (rfu). Previously analysed data sets provided by the laboratories were analysed with a maximum analysis threshold of 30 rfu, with some examples at thresholds below this. All profiles were amplified at 30 cycles (as per the manufacturer's recommendations). Amplified products from two laboratories were separated using Applied Biosystem's 3500 capillary electrophoresis instruments with the remaining laboratories using Applied Biosystems 3130 instruments.

A total of 1295 profiles were available from single source samples prepared at optimal conditions from pristine DNA. These profiles demonstrated a range of degradation slopes despite being pristine DNA at optimal amplification conditions. Fifty of the most degraded profiles were selected by taking those with the biggest difference in the ratio of the peak heights at Penta D (a high molecular weight locus) to D16S539 (a low molecular weight locus). These represent some of the profiles with the steepest downward slopes, and thus would be described as degraded using common terminology, whether caused by degradation or other phenomena such as inhibition or preferential amplification. As such, these profiles are more likely to provide the desired information on the nature of the relationship between peak height and molecular weight.

One consequence of a linear model is that there exists the possibility that predicted peak heights are negative. This is unreasonable. To avoid this possibility the linear function was modified as shown in equation (1). We write the expected peak height as E_a.

$$E'_a = \begin{cases} E_{a,} & E_a \geq Z/2 \\ Z/2, & E_a < Z/2 \end{cases} \quad (1)$$

where Z is the analytical threshold.

The models of interest are linear,

$$E'_a = \max(Z/2,\ t + d \times m_a) \quad (2)$$

and exponential

$$E'_a = t \times e^{d \times m_a} \quad (3)$$

where: t is the intercept of the line or the constant of proportionality for expected height vs. molecular weight; d is the slope of the line or exponent for expected height vs. molecular weight.

The values for t and d for each model were determined using maximum likelihood estimation in MS Excel. The exponential and linear models (equations (2) and (3)) were then fitted for each profile by least squares in MS Excel.

Results and conclusions

Figure 1 shows a plot of $\log(O_a/E'_a)$ versus molecular weight in base pairs using the exponential and modified linear models. We can see that more extreme positive departures from expectation occur at the high molecular weight end, approximately 350bp and above, and extreme negative departures occur in the mid-zone. This is expected if we force a straight line on an exponential curve. The cluster of high x data points at the right hand end of Figure 1 represent high molecular weight peaks that are observed but which are predicted to be absent using the linear model. The exponential fit reduces the number of these types of departure. The R^2 values are 0.82 and 0.86 for

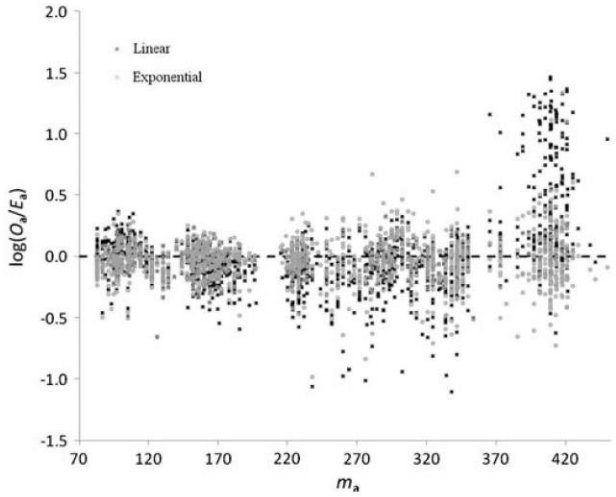


Figure 1. A plot of $\log(O_a/E'_a)$ versus m_a in base pairs using the exponential and linear fitting.

linear and exponential, respectively. The two models have the same number of free variables and hence any reduction in R^2 represents an improvement. We conclude that this evidence supports the use of an exponential curve to model peak heights versus molecular weight in PowerPlex 21 profiles.

Acknowledgements

We warmly acknowledge the comments of Catherine McGovern and Stuart Cooper and two anonymous reviewers that have greatly improved this paper. We acknowledge the gracious provision of data from Victoria Police Forensic Services Department. This work was supported in part by grant 2011-DN-BX-K541 from the US National Institute of Justice. Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the US Department of Justice.

References

1. Edwards A, Civitello A, Hammond HA, Caskey CT. DNA typing and genetic mapping with trimeric and tetrameric tandem repeats. Am J Hum Genet. 1991;49(4):746–756.
2. Ballantyne J, Hanson EK, Perlin MW. DNA mixture genotyping by probabilistic computer interpretation of binomially-sampled laser captured cell populations: combining quantitative data for greater identification information. Sci Justice. In press, doi:10.1016/j.scijus.2012.04.004.
3. Bill M, Gill P, Curran J, Clayton T, Pinchin R, Healy M, et al. PENDULUM—a guideline-based approach to the interpretation of STR mixtures. Forensic Sci Int. 2005;148(2–3):181–189.
4. Kelly H, Bright J-A, Curran JM, Buckleton J. Modelling heterozygote balance in forensic DNA profiles. Forensic Sci Int: Genet. 2012;6(6):729–734.
5. Nicklas JA, Noreault-Conti T, Buel E. Development of a real-time method to detect DNA degradation in forensic samples. J Forensic Sci. 2012;57(2):466–471.
6. Chung DT, Drábek J, Opel KL, Butler JM, McCord BR. A study on the effects of degradation and template concentration on the amplification efficiency of the STR Miniplex primer sets. J Forensic Sci. 2004;49(4):733–740.
7. McCord B, Opel K, Funes M, Zoppis S, Jantz LM. An investigation of the effect of DNA degradation and inhibition on PCR amplification of single source and mixed forensic samples. 2011, available from: https://www.ncjrs.gov/App/Publications/abstract.aspx?ID=258707
8. Cotton EA, Allsop RF, Guest JL, Frazier RRE, Koumi P, Callow IP, et al. Validation of the AMPF*l*STR SGM Plus system for use in forensic casework. Forensic Sci Int. 2000;112(2–3):151–161.
9. Diegoli TM, Farr M, Cromartie C, Coble MD, Bille TW. An optimized protocol for forensic application of the PreCR Repair Mix to multiplex STR amplification of UV-damaged DNA. Forensic Sci Int: Genet. 2012;6(4):498–503.
10. Bright JA, Taylor D, Curran JM, Buckleton JS. A biological model for a continuous method of DNA interpretation. In press, 2013, http://dx.doi.org/10.1016/j.fsigen.2012.11.013
11. Buckleton J, Triggs CM. Is the 2p rule always conservative? Forensic Sci Int. 2006;159:206–209.
12. Buckleton JS, Triggs CM, Walsh SJ. Forensic DNA evidence interpretation. Boca Raton (Florida): CRC Press; 2004.
13. Haned H. Forensim: an open-source initiative for the evaluation of statistical methods in forensic genetics. Forensic Sci Int: Genet. 2011;5(4):265–268.
14. Haned H, Gill P. Analysis of complex DNA mixtures using the Forensim package. Forensic Sci Int: Genet Suppl Series. 2011;3(1):e79–e80.

15. Balding DJ, Buckleton J. Interpreting low template DNA profiles. Forensic Sci Int: Genet. 2009;4(1):1–10.
16. Lohmueller KE, Rudin N. Calculating the weight of evidence in low-template forensic DNA casework. J Forensic Sci. 2012;n/a–n/a.
17. Perlin MW, Legler MM, Spencer CE, Smith JL, Allan WP, Belrose JL, et al. Validating TrueAllele DNA mixture interpretation. J Forensic Sci. 2011;56(6):1430–1447.
18. Evett IW, Gill PD, Lambert JA. Taking account of peak areas when interpreting mixed DNA profiles. J Forensic Sci. 1998;43(1):62–69.
19. Geddes L. What are the chances? New Scientist. 2010;207(2774):8–10.
20. Geddes L, King C, Stier C. Between prison and freedom. New Scientist. 2010;207(2773):8–11.
21. Tvedebrink T, Eriksen PS, Mogensen HS, Morling N. Statistical model for degraded DNA samples and adjusted probabilities for allelic drop-out. Forensic Sci Int: Genet. 2012;6(1):97–101.

2.4: Drop out

Manuscript: Utilising allelic dropout probabilities estimated by logistic regression in casework. J Buckleton, H Kelly, JA Bright, D Taylor, T Tvedebrink, JM Curran. (2014) Forensic Science International: Genetics 9, 9-11 – *Cited 9 times*

Statement of novelty: This work provides a comparison of the performance of several published methods for modelling drop-out and a new variant developed and described in the paper.

My contribution: I was a minor contributor to the modelling work carried out and assisted in writing the manuscript (in numerical terms approximately 20%).

Research Design / Data Collection / Writing and Editing = 10% / 5% / 20%

Additional comments: The models of drop-out developed in this paper are not those that were ultimately used in STRmix™. This paper is included to complete the picture of DNA profile behaviour modelling that is the focus of chapter 2

Forensic Science International: Genetics 9 (2014) 9–11

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

Utilising allelic dropout probabilities estimated by logistic regression in casework

John Buckleton [a,*], Hannah Kelly [b], Jo-Anne Bright [a,b], Duncan Taylor [c], Torben Tvedebrink [d,e], James M. Curran [b]

[a] ESR Ltd, Private Bag 92021, Auckland, New Zealand
[b] Department of Statistics, University of Auckland, Private Bag 92019, Auckland, New Zealand
[c] Forensic Science South Australia, 21 Divett Place, SA 5000, Australia
[d] Department of Mathematical Sciences, Aalborg University, Fredrik Bajers Vej 7G, DK-9220 Aalborg East, Denmark
[e] Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health Sciences, University of Copenhagen, Denmark

ARTICLE INFO

Article history:
Received 15 January 2013
Received in revised form 30 June 2013
Accepted 1 July 2013

Keywords:
Forensic DNA interpretation
Dropout
Logistic regression
Low template

ABSTRACT

Some advanced methods for DNA profile interpretation require a probability for the event of dropout. Methods have been suggested based on logistic regression. Two of these respectively use a proxy for template that is constant across loci and one that is modelled using an exponential curve. Both of these methods allow different modelling constants from each locus. A variant of the model using an exponential curve is discussed. This variant constrains the constants to be the same for every locus. We test these two methods and the variant by developing the constants (training) on one set of data and testing them on another. This mimics the likely use in casework. We find that the new variant appears to be the most useful in that it performs better than the other two options when trained on one data set and used on another. The hypothesised reason for this is that locus to locus variation in amplification efficiency varies with time, multimix batch, or from sample to sample.

1. Introduction

In forensic DNA analysis, routine casework is often undertaken using short tandem repeat (STR) loci amplified by the polymerase chain reaction (PCR). If the sample is low level or degraded some alleles from the true DNA donors may fail to produce a peak above threshold in the resulting electropherogram (epg). This phenomenon is termed allelic dropout or simply dropout.

Tvedebrink et al. and Gill et al. introduced the concept of modelling the probability of dropout using logistic regression [1,2]. Logistic regression is a type of regression analysis that can be used to predict the probability[1] of a binary event, based on one or more explanatory variables. In the context of a DNA profile the best explanatory variable to predict the probability of dropout would be the true, but unknown, template available at each locus for amplification. This introduces the concept that the available template at each locus differs.

Often epgs arising from casework exhibit a decrease in allelic peak height as the molecular weight (w) of the alleles increase. This is variously described as the degradation slope or the 'ski slope' [3,4]. Effective modelling of degradation is likely to provide the most effective explanatory variable for dropout.

2. Method

In the following method section we test several plausible models. Following Tvedebrink et al. [1,5,6] we could envisage that $\hat{H}$ serves as a proxy for template pre-amplification and is thought to be constant at every allelic position. We introduce the term mass, and denote it $\hat{H}_a$, to subsume the concepts of template number and degradation. Hence $\hat{H}_a$ serves as the proxy at allelic position a.

If the degradation of the DNA strand was random with respect to location, then we would anticipate that the expected height of peak a, E_a, would be exponentially related to molecular weight, w_a, and to whether the peak was a heterozygote or homozygote. Let X_a be the count of allele a. $X_a = 1$ for a heterozygote with a and $X_a = 2$ for a homozygote a. The expected height, E_a, of peak a is therefore modelled as

$$E_a = \hat{H}_a X_a$$
$$\hat{H}_a = \alpha_0 e^{\alpha_1 w_a}$$

* Corresponding author. Tel.: +64 9 8153 904; fax: +64 9 8496 046.
E-mail address: john.buckleton@esr.cri.nz (J. Buckleton).

[1] Actually the logarithm of the odds, but there is a simple transformation between probability and log odds.

1872-4973/$ – see front matter © 2013 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.fsigen.2013.07.001

A decreasing exponential relationship between allele height and molecular weight was described by Tvedebrink et al. [2] in relation to models for allelic dropout and has been confirmed at least once empirically.

Experience in casework has also suggested that there is a locus effect in addition to a general downward slope. As an example, in one report three loci within the Identifiler™ mulitplex were preferentially inhibited to varying extents in the presence of a laboratory cleaning agent [3]. Multimix is produced in batches and it is conceivable that the locus balance in one batch is different from another. Equally inhibitors co-extracted with the sample could affect certain loci more than others. The cameras used to detect the fluorescence have been shown to differ in their response to the different dyes used in detection and it is likely that the camera response changes as the camera ages. Collectively these factors suggest that loci may be above or below the trendline and that whether a specific locus is above or below may change from time to time or even sample to sample.

Models ignorant of such effects are likely to underperform.

We will discuss three models for the probability of dropout of a single allele D_a and of a homozygote, D_{2a}, as

$$D_a = \frac{e^{(l_{i0}+\beta_0)+(l_{i1}+\beta_1)\ln \hat{H}}}{1+e^{(l_{i0}+\beta_0)+(l_{i1}+\beta_1)\ln \hat{H}}} \quad D_{2a}$$

$$= \frac{e^{(l_{i0}+\beta_0)+(l_{i1}+\beta_1)\ln 2\hat{H}}}{1+e^{(l_{i0}+\beta_0)+(l_{i1}+\beta_1)\ln 2\hat{H}}} \quad (\text{for locus } i)\ldots \mathrm{T}_1 \text{ model}$$

$$D_a = \frac{e^{(l_{i0}+\beta_0)+(l_{i1}+\beta_1)\ln \hat{H}_a}}{1+e^{(l_{i0}+\beta_0)+(l_{i1}+\beta_1)\ln \hat{H}_a}} \quad (\text{for locus } i)\ldots \mathrm{T}_2 \text{ model}$$

$$D_a = \frac{e^{\beta_0+\beta_1 \ln \hat{H}_a}}{1+e^{\beta_0+\beta_1 \ln \hat{H}_a}} \quad (\text{for locus } i)\ldots \mathrm{T}_2' \text{ model}$$

There are currently two published logistic regression dropout models.

The first model (T_1), published by Tvedebrink et al. in 2009 [3], uses an average of peak heights across a full profile, $\hat{H}$, as a proxy for mass. A logistic model was fitted allowing separate logistic parameters β_0 and β_1 for each locus

Tvedebrink et al. [7] subsequently published a second model (T_2) that, correctly, models mass as an exponential function of molecular weight, $\hat{H}_a$. This model also allows separate β_0 and β_1 for each locus.

Therefore in model T_1 one DNA proxy is used for all loci whereas in T_2 the DNA proxy variable is locus-dependent.

However the question presents, do locus effects developed for one set of 'training' data translate to a future set of data? This question is more than academic. If multimix batches or even samples differ in locus amplification efficiency then transportability of the model to future profiles may be an issue. Accordingly it may be advantageous to consider a model that incorporates the concept of degradation but does not include a locus effect.

We depart from the much tidier terminology of Tvedebrink et al. [1,3] by the use of D_a for Pr(D) for a heterozygote with allele a and D_{2a} for a homozygote aa to align with our other work.

Tvedebrink et al. [1] initially used a function of $\ln \hat{H}$ as the explanatory variable (T_1 model). For the exponential model Tvedebrink et al. [3] utilised $\ln \hat{H}_a$ (T_2 model). In this work a model using $\ln \hat{H}_a$ is trialled without a locus effect (T_2' model).

The constants in the above models are developed from empirical data by logistical regression. In T_1 and T_2 the values, l_{i0} and l_{i1}, vary across loci. In T_2' one β_0 and one β_1 are applied to all loci. Note that in model T_1 the subscript a could be dropped as the dropout probabilities apply to the entire profile and are not allele specific, however we retain them here for consistency.

In practical application it is likely that the constants will be developed on one set of data (termed the training set here) and then applied to casework. In order to assess the suitability of the models when used in this way we develop the constants on the training set and then apply them to a different set of data termed the test set.

The three models described (T_1, T_2 and T_2') were applied to the datasets outlined below.

Case files were examined and data collated for profiles suggesting a single contributor where the circumstances allowed a reasonable inference about the source. The case file dates varied from November 2009 to May 2012. The DNA samples had been extracted using DNA IQ™ (Promega Corporation) method for saliva, bloodstains and trace samples. All samples were quantified using Applied Biosystems Quantifiler™ human DNA detection system (Life Technologies, Carlsbad CA) and 1.5 ng of DNA was targeted for Applied Biosystems Identifiler™ (Life Technologies, Carlsbad CA) amplification on a 9700 thermal cycler (Applied Biosystems) with a silver block. Amplified DNA was analysed using a 3130xl capillary electrophoresis instrument and DNA profile data was analysed using GeneMapper™ ID software (Applied Biosystems). These data, total 213, were split into three equal parts and termed I_1, I_2 and I_3.

120 single source profiles of pristine DNA of known origin from blood and semen stains were obtained. Samples were analysed as for the sets above. This dataset was termed I_4.

Pristine DNA from buccal swabs collected from 10 volunteers were extracted using DNA IQ™ (Promega) as per the manufacturer's directions. Extracted DNA was quantified twice using Applied Biosystems Quantifiler™ human DNA detection system (Life Technologies, Carlsbad CA) and an average taken. Varying quantities of DNA (1 ng, 500 pg, 250 pg, 100 pg, 75 pg, 50 pg, 10 pg, 5 pg and 1 pg) were amplified using Promega's PowerPlex® 21 System in 12.5 μL reactions on a 9700 thermal cycler (Applied Biosystems) with a silver block. Amplified DNA was analysed using a 3130xl capillary electrophoresis instrument and DNA profile data was analysed using GeneMapper™ ID software (Applied Biosystems). This produced 70 data termed set P.

There is some interest in the total number of free variables. This arises because a significant factor when considering the future performance of a model is whether overfitting has occurred on the training set. Overfitting occurs when there are sufficient free parameters to allow fitting to some aspects of training set that do not appear in future test sets or casework. T_1 allows separate logistic parameters β_0 and β_1 for each locus giving a total of $2l$ parameters for l loci. There is one covariate, the proxy for template, $\hat{H}$ extracted from each profile. Plausibly total parameters are $2l + n$. T_2 allows separate β_0 and β_1 for each locus. There are two covariates per profile: the two parameters of the exponential curve. Plausibly total parameters are $2l + 2n$. For our proposed model T_2' there are $2n + 2$ free parameters.

$\hat{H}$ and $\hat{H}_a$ values were both obtained using least squares fitting. For the methods T_1 and T_2 the l_{i0} values were constrained to within a factor of two of the average. This avoided them moving to unreasonable values. For the various sets, one was chosen to train the approach which was then tested against the others. The threshold for dropout was set at 50 RFU. Peaks above this were deemed present and those below this deemed absent.

3. Results

We use the mean log(likelihood) per profile to score each of the models. The model which yields highest mean log(likelihood) is regarded as the preferred model. Therefore, if on average one model yields higher log(likelihood) values than the other models,

Table 1
Mean log(likelihood) for different combinations of model and training sets $I_1, \ldots, I_4$ and I'_1 and I'_2. The set used for training the parameters is marked – T. The highest value in each row is marked in bold.

Data set	Method applied		
	T_1	T_2	T'_2
I_1 – T	**−0.20**	**−0.20**	−0.34
I_2	−3.84	−3.66	**−2.00**
I_3	−4.51	−4.06	**−2.03**
I_4	−2.79	−2.61	**−0.98**
I_1	−0.46	−0.45	**−0.42**
I_2 – T	−1.31	**−1.30**	−1.77
I_3	−2.21	−2.16	**−1.91**
I_4	−1.17	−1.14	**−1.08**
I_1	−0.46	−0.43	**−0.35**
I_2	**−1.64**	−1.74	−1.82
I_3 – T	−1.79	**−1.71**	−1.91
I_4	−1.29	−1.26	**−1.02**
I_1	−0.44	−0.44	**−0.34**
I_2	−2.02	−2.00	**−1.98**
I_3	−2.61	−2.39	**−2.02**
I_4 – T	**−0.82**	**−0.82**	−0.97

Table 2
Mean log(likelihood) for different splits of model and training sets for the PowerPlex® 21 set. Set P was pristine DNA. Five different 35:35 splits of the same data were trialled with one half used to train and one to test the approaches. The set used for training the parameters is marked – T. The highest value in each row is marked in bold.

	Method applied		
	T_1	T_2	T'_2
T	**−1.6**	−1.7	−2.2
	−3.2	−3.3	−3.8
T	**−1.7**	−1.8	−2.8
	−3.3	−3.4	**−3.2**
T	**−2.2**	−2.3	−2.9
	−2.6	−2.7	−3.1
T	**−1.6**	−1.7	−2.4
	−3.5	−3.6	−3.6
T	**−2.6**	−2.7	−3.6
	−2.1	**−2.1**	−2.4

then we would regard this as evidence of superior performance. The results are given in Tables 1 and 2.

4. Discussion

In almost all tests performed on the Identifiler™ sets method T_1 or T_2 produced the highest mean log(likelihood) in the training set and T'_2 produced the highest mean log(likelihood) in the test sets. The mean in the training set was always higher than in the test sets regardless of model used. We interpret this as meaning that a locus effect does exist but that this changes either batch to batch of mutlimix, or profile to profile, or over time by ageing of the camera or other laboratory changes such as cleaning agents. This means that a set of l_{i0} and l_{i1} values developed on one dataset, say during validation, is not necessarily transportable to new profiles developed subsequently.

For the PowerPlex® 21 data T_1 regularly gave the highest mean log(likelihood) in the training and test sets (Table 1). We interpret this as meaning that pristine source data is too good to show the expected degradation effect and therefore not suitable to train these logistic models.

Of the three methods studied T'_2 trained on casework data is narrowly the best for immediate use in casework due to its portability since it produced the highest log(likelihood) in test sets more often. However, we conclude that further development is required in the application of locus specific effects and it is likely that these locus effects will vary from profile to profile or time to time.

Acknowledgements

We gratefully acknowledge the valuable comments of Sue Vintiner, Johanna Veth, and two anonymous referees that have greatly improved this manuscript. This work was supported in part by grant 2011-DN-BX-K541 from the US National Institute of Justice. Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the U.S. Department of Justice.

References

[1] T. Tvedebrink, P.S. Eriksen, H.S. Mogensen, N. Morling, Estimating the probability of allelic drop-out of STR alleles in forensic genetics, Forensic Sci. Int. Genet. 3 (2009) 222–226.
[2] T. Tvedebrink, P.S. Eriksen, M. Asplund, H.S. Mogensen, N. Morling, Allelic drop-out probabilities estimated by logistic regression – further considerations and practical implementation, Forensic Sci. Int. Genet. 6 (2012) 263–267.
[3] T. Tvedebrink, P.S. Eriksen, H.S. Mogensen, N. Morling, Statistical model for degraded DNA samples and adjusted probabilities for allelic drop-out, Forensic Sci. Int. Genet. 6 (2012) 97–101.
[4] P. Gill, R. Puch-Solis, J. Curran, The low-template DNA (stochastic) threshold – its determination relative to risk analysis for national DNA databases, Forensic Sci. Int. Genet. 3 (2009) 104–111.
[5] J.A. Nicklas, T. Noreault-Conti, E. Buel, Development of a real-time method to detect DNA degradation in forensic samples, J. Forensic Sci. 57 (2012) 466–471.
[6] D.T. Chung, J. Drabek, K.L. Opel, J.M. Butler, B.R. McCord, A study of the effects of degradation and template concentration on the amplification efficiency of the STR Miniplex primer sets, J. Forensic Sci. 49 (2004) 733–740.
[7] J.A. Bright, S. Cockerton, S. Harbison, A. Russell, O. Samson, K. Stevenson, The effect of cleaning agents on the ability to obtain dna profiles using the Identifiler (TM) and PowerPlex (R) Y Multiplex Kits, J. Forensic Sci. 56 (2011) 181–185.

2.5: Saturation, baseline and drop-in

Manuscript: Validating multiplexes for use in conjunction with modern interpretation strategies. D Taylor, JA Bright, C McGovern, C Hefford, T Kalafut, J Buckleton. (2016) Forensic Science International: Genetics 20, 6-19 – *Cited 7 times*

Statement of novelty: The calibration of STRmix™ for a specific laboratories data requires that they examine a number of aspects of DNA profile behaviour. While papers and books in the past have commented on these behaviours, this paper provides a description of statistical models for each behaviour and the means of modelling them. Many of the models described in this work replace rules or thresholds that were typically used to interpret DNA profiles

My contribution: I was the main author and did the majority of modelling and analysis that went into this manuscript.

Research Design / Data Collection / Writing and Editing = 60% / 75% / 60%

Additional comments:

Forensic Science International: Genetics 20 (2016) 6–19

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

Research paper

Validating multiplexes for use in conjunction with modern interpretation strategies

Duncan Taylor[a,b,*], Jo-Anne Bright[c], Catherine McGoven[c], Christopher Hefford[a], Tim Kalafut[d], John Buckleton[c]

[a] Forensic Science South Australia, 21 Divett Place, Adelaide, SA 5000, Australia
[b] School of Biological Sciences, Flinders University, GPO Box 2100, Adelaide, SA 5001, Australia
[c] Institute of Environmental Science and Research Limited, Private Bag 92021, Auckland 1142 New Zealand
[d] U.S Army Criminal Investigation Laboratory, Defense Forensic Science Centre, Forest Park, GA 30297, USA

ARTICLE INFO

Article history:
Received 22 March 2015
Received in revised form 21 September 2015
Accepted 22 September 2015

Keywords:
Forensic DNA interpretation
GlobalFiler
Continuous DNA interpretation
STRmix
Modelling
DNA mixtures

ABSTRACT

In response to requests from the forensic community, commercial companies are generating larger, more sensitive, and more discriminating STR multiplexes. These multiplexes are now applied to a wider range of samples including complex multi-person mixtures. In parallel there is an overdue reappraisal of profile interpretation methodology. Aspects of this reappraisal include

1. The need for a quantitative understanding of allele and stutter peak heights and their variability,
2. An interest in reassessing the utility of smaller peaks below the often used analytical threshold,
3. A need to understand not just the occurrence of peak drop-in but also the height distribution of such peaks, and
4. A need to understand the limitations of the multiplex-interpretation strategy pair implemented.

In this work we present a full scheme for validation of a new multiplex that is suitable for informing modern interpretation practice. We predominantly use GlobalFiler™ as an example multiplex but we suggest that the aspects investigated here are fundamental to introducing any multiplex in the modern interpretation environment.

1. Introduction

In response to requests from the forensic community, commercial companies are generating larger, more sensitive, and more discriminating STR multiplexes. For example in 2010, the CODIS Core Loci Working Group was formed to investigate the expansion of the minimum load criteria to CODIS from 13 STR loci. One of the aims was to balance the total number of loci recommended with the level of discrimination offered in order to reduce the likelihood of adventitious matches and in anticipation of more transnational sharing of DNA profile information [1]. The gender determining locus Amelogenin, 18 autosomal STRs and one Y STR are the new minimum recommended STR marker set with another three autosomal STRs strongly recommended [1,2]. In Europe a core of 15 STRs has been designated as the European Standard Set [3].

These multiplexes are now applied to a wider range of samples including complex multi-person mixtures.

In parallel there is an overdue reappraisal of profile interpretation methodology. Aspects of this reappraisal include

1. The need for a quantitative understanding of allele and stutter peak heights and their variability,
2. An interest in reassessing the utility of smaller peaks below the often used analytical threshold,
3. A need to understand not just the occurrence of peak drop-in but also the height distribution of such peaks, and
4. A need to understand the limitations of the multiplex-interpretation strategy pair implemented.

In this paper we will outline a scheme for the validation of a multiplex that is suitable for use with modern interpretation strategies such as the semi and fully continuous systems being implemented in many parts of the forensic community. We emphasize that it is the multiplex-interpretation method couplet that requires validation. Hence slightly different suggestions might

* Corresponding author at: Forensic Science South Australia, 21 Divett Place, Adelaide, SA 5000, Australia. Fax: +61 8 8226 7777.
E-mail address: Duncan.Taylor@sa.gov.au (D. Taylor).

http://dx.doi.org/10.1016/j.fsigen.2015.09.011
1872-4973/

result for the same multiplex with a semi-continuous model than with a continuous model. We differentiate between developmental validation and internal validation. Developmental validation in this paper means actions we suggest should be undertaken by the software developer to ensure that the software is suitable for use on a certain multiplex.

The aspects we suggest should be studied are:

1. Noise,
2. Stutter ratio and variability,
3. Peak height variability both at and between loci,
4. Drop-out,
5. Drop-in, and
6. Saturation of the capillary electrophoresis (CE) camera.

We illustrate this scheme using the GlobalFiler™ and MiniFiler™ multiplexes. The GlobalFiler™ multiplex (Life Technologies, Carlsbad CA) amplifies 22 STRs, the gender marker Amelogenin plus an additional Y-indel locus [4]. The MiniFiler™ multiplex amplifies 8 autosomal STRs plus Amelogenin.

We finish the paper with sections regarding training and general legal acceptance of continuous and probabilistic DNA interpretation systems. While neither of these topics relates directly to kit validation within the laboratory (or at the computer) they remain an important part of any validation and are required before any pairing of software, expert and profiling kit can be introduced and defended in court.

2. Results

2.1. Analytical threshold

The change to probabilistic systems invites a reappraisal of our approach to setting the analytical threshold (AT). This is because modern systems can manage low level peaks better.

Two interpretation strategies are available:

1. A threshold (AT) based approach and
2. Systems that deal with potential noise at the interpretation stage and require no AT.

A peak in the electropherogram (epg) may be allelic, a PCR by-product, artefactual such as pull-up, or electronic noise. Back stutter is almost unavoidable and we can assume that almost every allelic peak has an associated back stutter peak. Forward stutter and double back stutter are also produced by the PCR process, but in smaller amounts. Since back stutter, forward stutter and double back stutter are allelic products they do not differ from a true allelic peak in any way and cannot be differentiated by visual examination. These are not the only artefactual PCR products. For example there is a −2 base pair stutter-like product at SE33 which is a complex locus with largely tetranucleotide repeats.

Discussions of the position of the AT usually concentrate on the electronic noise and it is suggested that the AT should not be used to manage artefacts. For example SWGDAM [5] states:

> . . . the analytical threshold should be established based on signal-to-noise considerations (i.e., distinguishing potential allelic peaks from background). The analytical threshold should not be established for purposes of avoiding artifact labeling as such may result in the potential loss of allelic data.

Valid efforts have been made to model electronic noise and we give a feel of these types of efforts in Appendix 1. These approaches usually consider the probability of a peak of height O_a if it is electronic noise, Pr(O_a|electronic noise). They suggest selecting an AT at some point when Pr(O_a|electronic noise) is expected to be small. We embrace the validity of the sentiment about not using an AT to manage artefacts but in a brutally pragmatic sense it is necessary to consider the downstream effects of the position of the AT. This needs a lot more than a consideration of Pr(O_a|electronic noise) and we suggest that electronic noise is the least difficult of the factors needing consideration. At the time of writing none of the probabilistic systems specifically model forward and double backward stutter. The semi-continuous systems do not model back stutter whereas the continuous ones do.

Consider initially a threshold based approach. Peaks above the AT are often examined manually for morphology. At this stage dye blobs, pull-up and electronic spikes will be removed. Any peak above the AT that passes manual inspection is passed to the interpretation phase. Lowering the AT will detect more allelic peaks but will also pass more artefactual peaks for manual inspection. The total utility of a lowering of the AT is therefore the sum of these effects and depends crucially on how significant the consequences of passing such peaks are. In turn, this depends on how they are treated at the interpretation phase.

In previous binary systems the passing of false peaks had a very significant negative effect (negative utility) on the interpretation. Hence, historically, ATs were set high. The modern systems have a greater resilience to false peaks and hence the utility function is changed. More specifically there is now less risk associated with lower ATs as long as the models within the system for DNA profile behaviors (such as drop-in) are set up accordingly.

The semi-continuous models in widespread use (LRmix [6], Lab Retriever, LikeLTD [7], FST [8] and LiRa [9]) do not utilise peak heights directly in the software. Some interaction of peak height data and semi-continuous systems does exist, for example expert intervention allows for the manual extraction of a clear major [10] and Lab Retriever utilizes peak height in forming the probability of drop-out (D) parameter. All three of these systems currently function with a threshold based strategy and peaks in stutter positions are either removed or dealt with as ambiguous (either partly allelic or totally stutter). Any peaks that are above the AT and passed to the software must be explained as allelic, ambiguous or as drop-in. Peaks dealt with using the drop-in function would include true drop-ins, that is, unreproducible allelic peaks that appear in the profile, and non-allelic peaks not treated as ambiguous. To emphasize that the software should be run by an expert (following Gill and Haned [6]) we will refer to the software-expert pair (SEP). The effect of dropping the AT would be more true allelic peaks detected which has a strongly positive utility. However there will also be more peaks needing manual removal, treatment as ambiguous or drop-in, or which may cause the scientists to artificially increase the assigned number of contributors. These have a negative utility. The net effect is unstudied.

The continuous software programmes in current use (STRmix™ [11] and TrueAllele [12]) can treat noise peaks directly; modelling the probability of these peaks if they arise from a contributor (as allelic or stutter) or if they do not arise from a contributor (encompassing noise or drop-in).

If a drop-in function is used for semi-continuous SEP or STRmix™ the AT value does not need to be set as conservatively as with traditional interpretation methods. The setting of an AT will affect both the probability of drop-out and drop-in. We direct the reader to [13] who explore this idea.

If forward stutter and double back stutter are not manually removed or treated as ambiguous then the drop-in function will now be used to model true drop-ins but also other peaks that pass AT and manual inspection. The drop-in rate will therefore need to be higher than if it was simply modelling drop-in. When used in this way the drop-in rate cannot be set from empirical negative control data but needs to be set from positive samples with known ground truths. In doing this there would be a dependence of profile

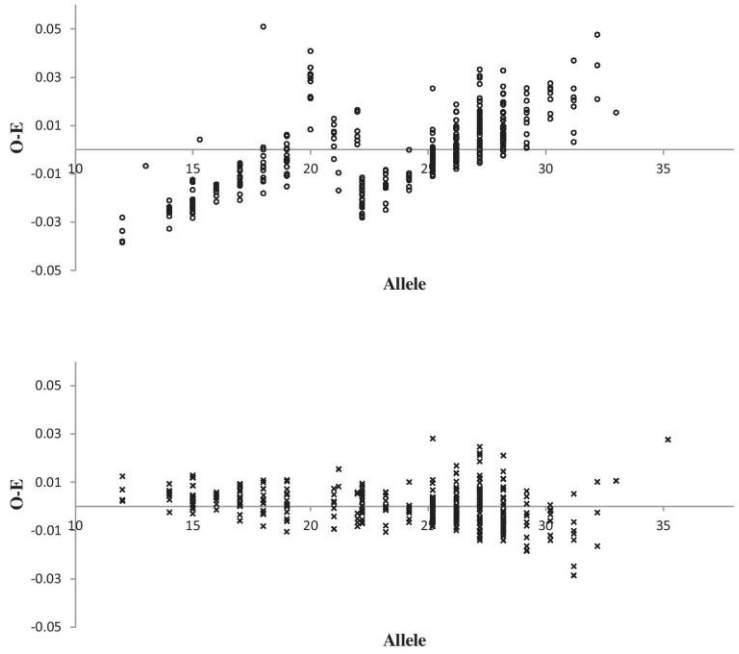


Fig. 1. Observed SR – the predicted SR (O–E) vs allele for the SE33 locus. LUS model (top) multi-sequence model (bottom).

intensity on drop-in (both in the validation data being used to determine drop-in rate, and in the evidential data that is later analyzed) that would be ignored.

It is therefore advantageous for the semi-continuous SEP and STRmix™ to manually remove forward and double back stutter until it can be modeled appropriately. We discuss forward stutter in a later section but at this stage cannot add to the discussion on double back stutter.

2.2. Back stutter ratio (SR)

Stutter ratio is quantified by calculating the ratio of the observed height of the stutter peak (O_{a-1}) to the observed height of the allelic peak (O_a). Stutter has been extensively modelled in the literature and was initially reported that the longest uninterrupted repeat sequence (LUS) was the best indicator of stutter ratio [14–17].

SE33 stutter ratios are poorly explained by a model based on LUS[1]. This locus has some alleles with two or three long sequences. The locus SE33 better fits a model where all repeat sequences are considered as contributing to stuttering but only after subtraction of a factor, x, of repeats. We term this the multi-sequence model. When considering the multi-sequence model we fit the equation:

$$\mathrm{SR} = m\sum_i \max(l_i - x, 0) \tag{1a}$$

where m and x are constants set by least squares. The residuals when applying the model specified in Eq. (1a) (and later (1b)) are normal distributed. l_i is the length of sequence i. The original LUS model was $\mathrm{SR} = ml_1 + c$ where m and c were constants and l_1 was the length of the longest uninterrupted sequence. By constraining the multi-sequence model in this way the number of free variables in the two models is the same and hence a direct comparison can be made. We demonstrate this comparison with 205 GlobalFiler™ profiles. For some alleles there are up to four known sequences in STRbase. For these we have simply averaged the sequences given. The effect of this is a broadening of the distributions about the predicted SR.

In Fig. 1 is shown a plot of the observed SR minus the predicted SR ($O - E$) by either the LUS model (top) or by the multi-sequence model (bottom) versus allele designation for the SE33 locus.

The term x in the multi-sequence model can be interpreted as the number of repeats before stuttering begins. For SE33 this was $x = 5.83$, as determined by maximum likelihood estimation (MLE). We will term this the lag. For SE33 this means that the short sequences contribute nothing to stuttering and that only sequences longer than 6 repeats contribute at all. For SE33 the average error in estimation of SR was 8.7×10^{-4} for the LUS model and 4.4×10^{-4} for the multi-sequence model.

However the first dataset we examined appeared to be the best one for this model. In other datasets we have found it necessary to modify the multi-sequence model to

$$\mathrm{SR} = m\sum_i \max(l_i - x, 0) + c \qquad \text{where } c \text{ is a constant.} \tag{1b}$$

Using this amendment the multi-sequence model can therefore completely replace the LUS model for all loci studied to date. We provide as supplementary material an Excel spreadsheet that demonstrates (with instructions) the application of this model to SE33 data.

[1] We thank Shawn Montpetit and Melissa Strong of the San Diego Police Department for bringing this to our attention.

The multi-sequence theory has a better intuitive feel, to us, than the LUS theory. However if it has validity it should be able to explain SR at least as well as LUS for the other loci. The formula for the LUS model can be rewritten into the same form as the multi-sequence model$SR = m(l_1 + \frac{c}{m})$ where c is typically negative. It will return the same result as the multi-sequence result whenever there is only one sequence per allele longer than $-(c/m)$ (note that when c is positive this term will be negative, meaning that any length sequence will contribute to the stutter). For the simple repeat loci there is no difference between allele, LUS and a multi-sequence approach.

If we consider a locus such as TH01, there is one common allele, the 9.3, which has a sequence interruption. The sequence contains a stretch of 6 repeat units and one incomplete unit of 3 bp. Hence a 9.3 allele has a LUS of 6. The multi-sequence model returns $x = 3.32$. This means that the first 3 repeat sequences contribute nothing to the modelled SR. Hence, for TH01, there is no difference between the LUS and the multi-sequence models. This is also true for FGA where no secondary sequence exceeds the value for x of 4.80. For vWA and D21S11 some secondary and tertiary sequences are of a moderate length. For these two loci the multi-sequence model showed a reduction in the average error in estimation of 17% and 5% respectively compared with the LUS model. This is consistent with the small effect of secondary and tertiary sequences above the lags of $x = 3.46$ and 4.66 respectively.

For the internal validation of a new multiplex we recommend that mean and variability of SR should be determined for a range of genotypes (an arbitrary number that has, in the author's experience, been adequate is 100). Single source samples are acceptable and should vary in template. Epgs should be analysed to very low peak height values, say 10 RFU, regardless what AT is to be used in this SEP. Ignore loci where the high molecular weight allele can stutter onto the low molecular weight allele (i.e. stutter affected heterozygous genotypes).

SR should be fitted to $\sum_i \max(l_i - x, 0)$ to confirm the relationship and inform the model for each locus. Least squares fitting appears to perform well.

For SEPs using semi-continuous systems, peaks in stutter positions are currently designated as allelic, stutter or ambiguous. This is based on a stutter threshold. At the time of writing two threshold structures are in use. These are:

1. One threshold value for SR across all loci, or
2. A separate threshold value for SR for each locus.

Peaks above this threshold are deemed to be allelic. Peaks below this threshold are deemed to be stutter if there is no minor contributor approximately equal to their height in the profile or if all minor allelic peaks have been seen at this locus. If a peak is neither designated allelic nor stutter then it is called ambiguous.

Thresholds for stutter are typically determined as the maximum stutter ratio found in some set of data. However the data presented above show that this maximum will be most affected by alleles with high LUS values and low peak heights. For SEPs using this approach the experiments described above are not essential and represent an almost academic exercise to check that the multiplex is behaving as expected. Current threshold based approaches to stutter designation in use in the semi-continuous SEP are not aligned with the empirical observations and it is very hard to see how to improve them without a significant increase in the complexity of currently manual functions. The most probable impact of these inefficiencies is that a small amount of information present in the profile is unused in a fraction of cases.

The situation for the continuous SEPs is very different. These have the capability to utilise this information directly. Therefore, at least for STRmix™, these experiments or equivalents provide direct empirical information for use in the software.

2.3. Forward stutter ratio

Forward stutter ratio (FSR) is quantified by calculating the ratio of the observed height of the forward stutter peak (O_{a+1}) to the observed height of the allelic peak (O_a).

$$\text{FSR} = \frac{O_{a+1}}{O_a}$$

In most multiplexes forward stutter is observed rarely at autosomal loci with the exception of the one autosomal trinucleotide repeat marker in common use, D22S1045. Using standard casework and validation samples, explanatory variables previously used to predict back stutter height such as parent allele height (O_a), locus and the longest uninterrupted sequence (LUS) were found unsuitable for predicting the height of forward stutter peaks for all tetra and pentanucleotide repeats [18]. This result was surprising given the known mechanism of stutter generation and the anecdotal observations of forward stutter being observed commonly for parent peaks that had high peak heights (suggesting that there is indeed some dependence of forward stutter on parent peak height).

It is hypothesised here that forward stutter peak height does depend on parent peak height, however in [18], only the highest few percent of all forward stutter peaks were seen and this caused dependencies to be hidden. Fig. 2 shows that from all applicable forward stutter positions only a few percent had forward stutters observed (shown as the black, 'Original work' line). It was surmised that this was due to an AT of 30 RFU being applied to the original data analysis. To explore this idea a set of 55 single sourced, but

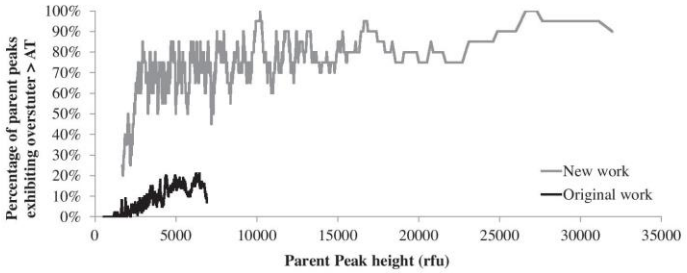


Fig. 2. Graph showing the percentage of parent peaks where forward stutter was observed (for genotypes where observation was possible i.e. not masked by other alleles or back stutter) for the original work carried out in [18] and current new work shown in this paper.

highly over saturated, GlobalFiler™ 3130xl generated profiles were analysed down to the level of 10 RFU. This lead to observed forward stutter peaks in more than 50% of instances (Fig. 2 shown as the grey, 'New work' line). In order to determine forward stutter ratio (FSR) the height of the parent peak had to be estimated, as observed parent peak heights could not be reliable upon due to oversaturation (see the saturation Section 2.6). The observed height of the back stutter (O_{a-1}) and the stutter ratio for that allele and locus (SR_a^l) were used to calculate the expected parent peak height (Ea) (Eq. (2)).

$$\mathrm{FSR} = \frac{O_{a+1}}{E_a} = \frac{\mathrm{SR}_a^l \times O_{a+1}}{O_{a-1}} \quad (2)$$

Consequently the parent peak height in Fig. 2 extends to 35,000 RFU while the 3130xl instrument saturates well below this intensity. The residuals when applying the model specified in Eq. (2) are normal distributed.

When the second allele of a heterozygous pair was one repeat unit different from the first allele, then the observed back stutter peak height of the first allele was used to determine expected parent peak height of the second allele. All instances where back stutter and forward stutter overlapped were omitted from the study. When a forward stutter was not observed then a value of 5 RFU (half the AT) was used.

Observed forward stutter peak height was regressed using R [19] with only expected parent peak height as a variable (forcing the intercept through the origin) and dependence was observed ($p < 2 \times 10^{-16}$, for the FSR being dependent on parent peak height). The regression was extended to include individual locus effects using the formula:

$$E_{a+1} = \beta_i E_a = \beta_i \left(\frac{O_{a-1}}{\mathrm{SR}_a^l} \right)$$

where a separate β_i term is present for each locus. The majority of loci showed a significant difference from each other and so locus is also a dependant variable for forward stutter peak heights. The values of β_i can be used to determine the FSR for each locus. The FSR and the standard errors from the regression are given below in Table 1, note that YSTR locus DYS391 has been omitted from the calculations, as has D22S1045 due to its strong correlation between LUS and FSR.

Table 1
regression results of forward stutter.

Locus	FSR	Std. error
CSF1PO	0.0062	0.001763
D10S1248	0.0128	0.001734
D12S391	0.0029	0.001539
D13S317	0.0046	0.001458
D16S539	0.0059	0.001536
D18S51	0.0045	0.001521
D19S433	0.0019	0.001739
D1S1656	0.0052	0.001533
D21S11	0.0072	0.001539
D2S1338	0.0016	0.001756
D2S441	0.0055	0.001695
D3S1358	0.0042	0.001624
D5S818	0.0077	0.001565
D7S820	0.0020	0.001518
D8S1179	0.0054	0.001524
FGA	0.0030	0.001566
SE33	0.0059	0.001575
TH01	0.0006	0.001613
TPOX	0.0007	0.001422
vWA	0.0033	0.000599

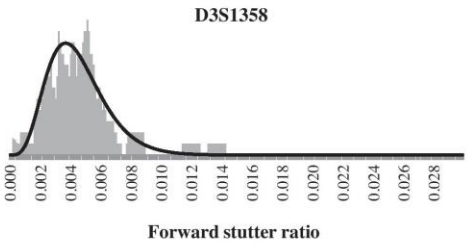


Fig. 3. Distribution of FSR for D3S1358 (grey) with a fitted gamma curve (black).

Note that it is likely that other factors will have an effect on forward stuttering such as the laboratory process, hardware or profiling system used.

2.3.1. Modelling forward stutter to inform threshold approaches

At time of writing we are aware of no SEP that directly models forward stutter peak heights. Therefore the requirement is to establish guidance on whether peaks in forward stutter positions are allelic. Due to the very low nature of forward stutter there is an inherent difficulty in accurate estimation. Many instances of forward stutter observed in these data were less than 30 RFU even for the oversaturated profiles examined. At this level the signal is mixed with the noise associated with the instrument. It is also in the range where small artefacts such as dye blobs, spikes, pull-up peaks or other exotic stutters (double back stutter, stutter-like artefacts etc.) from other alleles can all have large relative effects on peak height. Furthermore using back stutter peak heights to determine expected parent peak heights, which are then used to calculate the FSR, adds uncertainty to the estimation as there is stochastic variability in the back stutter process.

Compounding these difficulties is the fact that in standard validation studies or casework only those instances of forward stutter that are in the upper range are seen. This makes forward stutter modelling a difficult task.

Using the oversaturated data the distribution of FSR looks to have a long positive tail. This suggests that it could be modelled using a gamma distribution such as shown in Fig. 3 for D3S1358. In Fig. 3 the grey data shows the observations of FSR for D3S1358

Table 2
Gamma distributions fitted to FSR (shape and scale parameters rounded to four decimal places) and descriptive statistics of interest.

	Gamma parameters	Mode	Mean	95th quantile	99.9th quantile
D3S1358	5.6254, 0.0008	0.0038	0.0046	0.0081	0.0129
vWA	2.0089, 0.0018	0.0018	0.0036	0.0084	0.0164
D16S539	11.5371, 0.0006	0.0065	0.0071	0.0109	0.0154
CSF1PO	4.9010, 0.0013	0.0049	0.0062	0.0114	0.0185
TPOX	1.7981, 0.0005	0.0004	0.001	0.0024	0.0048
D8S1179	7.4452, 0.0008	0.0055	0.0063	0.0105	0.0159
D21S11	7.1822, 0.0011	0.0069	0.008	0.0134	0.0203
D18S51	3.1202, 0.0014	0.003	0.0044	0.0092	0.0162
D2S441	11.5938, 0.0006	0.0061	0.0066	0.0101	0.0143
D19S433	1.7754, 0.0009	0.0007	0.0016	0.004	0.0081
TH01	1.6952, 0.0008	0.0005	0.0013	0.0032	0.0065
FGA	3.2343, 0.0015	0.0032	0.0047	0.0097	0.017
D5S818	13.1417, 0.0006	0.0073	0.0079	0.0118	0.0163
D13S317	1.3800, 0.0028	0.0011	0.0039	0.0105	0.0222
D7S820	2.8112, 0.0013	0.0023	0.0036	0.0077	0.0138
SE33	3.7885, 0.0016	0.0045	0.006	0.0119	0.0202
D10S1248	5.6862, 0.0016	0.0074	0.009	0.0159	0.0252
D1S1656	4.9322, 0.0012	0.0049	0.0061	0.0112	0.0181
D12S391	1.3165, 0.0014	0.0004	0.0018	0.005	0.0107
D2S1338	1.5725, 0.0005	0.0003	0.0008	0.002	0.0041

(with some smoothing applied for visual clarity) and the black line shows the gamma distribution fitted by MLE. It is worth noting that using non-oversaturated data only the part of the distribution to the right of approximately 0.007 would be seen.

The distributions of FSR for each locus are given in Table 2 along with the mode, mean, 95th and 99.9th quantiles. These quantiles could be used by other laboratories using the same CE equipment and chemistry to inform their thresholds. For other combinations of CE and chemistry it is likely that these values are inappropriate.

2.3.2. Modelling forward stutter preparatory for probabilistic systems

Using the ratios from Table 1 and expected parent peak heights the expected forward stutter peak heights were calculated. Following the model of [11] we plot $\log_{10}(O_{a+1}/E_{a+1})$ against the expected forward stutter peak height (Fig. 4). The lower dashed line seen in Fig. 4 corresponds with the lower limit at which forward stutter peaks could be unobserved in the profiles using a AT of 10 RFU. Taking into account that the AT restricts the lower values that $\log_{10}(O_{a+1}/E_{a+1})$ can take (particularly noticeable for lower expected peak heights) the same relationship as described in [11] for allelic peaks appears to exist for forward stutter peaks:

$$\log_{10}\left(\frac{O_{a+1}}{E_{a+1}}\right) \sim N\left(0, \frac{c^2}{E_{a+1}}\right)$$

Using a value of $c^2 = 2$ the bounds shown in Fig. 4 cover 95% of the data.

2.3.3. Validation suggestions

Given the current lack of probabilistic systems incorporating a forward stutter model, the modelling described above is excessive for laboratories currently implementing probabilistic software. All that is required to handle these peaks is to set some sensible threshold above which a peak will be considered allelic. A sensible value for the data shown in Table 2 might be 2% of parent peak height, which would eliminate most instances of forward stutter at most loci. For D22S1045 a simple threshold would also be a sensible way forward although use of LUS as a predictor may be envisaged for the future. For peaks below this value a laboratory has three options:

1. Treat as ambiguous,
2. Leave in and use the drop-in function as a proxy to model the peak,
3. Remove the peak.

We are perfectly comfortable with the removal of the peak in this way and any suggestion of bias should be ameliorated by the use of a preset threshold. We recognize this as an interim suggestion pending the implementation of effective forward stutter models.

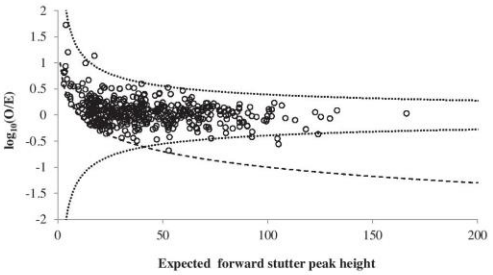


Fig. 4. Log(O/E) for forward stutter based on expected parent peak height showing 95% coverage using the model from [11] (dotted lines). The lower dashed line represents the limit at which data could be observed using an AT of 10 RFU.

In order to remove forward stutter peaks, some threshold is required and as stated previously the results in Table 2 may only apply to similar laboratory systems to that which was used to generate the data. It has also been stated that when only the upper tail of the forward stutter distribution is observed, modelling is complex and risks not revealing any dependencies. We suggest that to develop a threshold from standard validation data that a single, profile-wide, forward stutter threshold is used. The following information will be needed:

A – The proportion of applicable sites (locus and allele positions where forward stutter is not masked) where forward stutter was observed,

B – The proportion desired to be used as a cut-off for screening out forward stutter peaks, and

N – The number of observations of forward stutter.

When these values are known then FSR should be calculated for each N and listed in ascending order. The cut-off to be used is then the $N(A-1+B)$/Ath value in the list. For example if 500 observations of forward stutter are observed out of 5000 possible sites, then $A = 0.1$. The 500 observation as converted to FSR and listed in ascending order. If a cut-off is desired that removes 95% of all forward stutters then it should be set at the value of the:

$$500(0.1 - 1 + 0.95)/0.1 = 250\text{th}$$

value in the list. This very simple method makes the assumption that FSR are uniformly distributed above the AT, which is not true but as an approximation serves to produce a sensible cut-off value. More elegant modelling to produce a cut-off is possible but is more complex.

2.4. Peak height variability

Directly after the first use of fluorescence based STR profiling came the realisation that peak heights in DNA profiles are not reproducible. Partner peaks of heterozygous pairs appear at different heights (referred to as the heterozygous balance Hb or peak height ratio PHR), re-amplifications from the same DNA extract produce profiles with peaks that differ in height from each other, or fail to appear at all (referred to as drop-out).

For validation of a SEP we recommend analyzing a large set of single source samples of varying template. Ignore loci where the difference between the low and high molecular weight alleles is less than 2 repeat units (potential stutter affected heterozygotes). This is so that there is no overlap between the allele, stutters or forward stutters of a heterozygote pair. We introduce the concept here of total allelic product (TAP) [16] which describes the total amount of fluorescence expected from some number of strands of template DNA. During the PCR process replication errors occur that lead to stutter (either above and/or below the parent allele). Stutter strands also have had fluorescence incorporated, which would otherwise have been incorporated into the allelic peak. As expected by the TAP theory, it has been shown that higher than expected stutters are associated with a lower than expected allelic peaks [16]. We use the following terminology:

O_a – observed parent allele peak height.

O_{a-1} – observed stutter peak height.

O_{a+1} – observed forward stutter peak height.

O_{HMW} – observed TAP of the high molecular weight allele.

O_{LMW} – observed TAP of the low molecular weight allele.

Noting that for the two terms above $O_{(H/L)\text{MW}} = O_{a-1} + O_a + O_{a+1}$. We can then use O_{HMW} and O_{LMW} to calculate $\text{Hb}_{\text{TAP}} = \frac{O_{\text{HMW}}}{O_{\text{LMW}}}$ [20–23] and $\text{APH}_{\text{TAP}} = \frac{O_{\text{HMW}}+O_{\text{LMW}}}{2}$. We note that this resolution of detail is

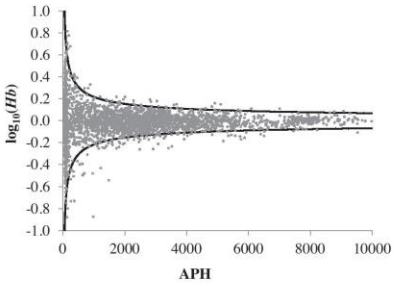


Fig. 5. Observed $\log_{10}$(Hb) data with 95% expected Hb boundaries from analysis of peak height variability

unlikely to substantially affect practical observations of empirical data, and for simplicity we retain the Hb and APH nomenclature. A plot of log(Hb) v APH is shown for the GlobalFiler™ data (grey points) in Fig. 5.

As might be expected from the nature of sampling variation, DNA extract sampling variation will have its largest relative effect on peak heights when the number of starting molecules of DNA is low. This effect has been shown in numerous graphs of Hb, measured over a range of average peak heights (APH) [16,22].

The peak height variability can be used to estimate the level of expected Hb in a dataset, as long as profiles within the dataset are not significantly affected by degradation (which will significantly affect the observed Hb variability, but not the peak height variability as long as degradation is included in the profile modelling). Taking advantage of the fact that a heterozygous peak pair comprises two peaks, each with height variability, the expected 95% bounds on expected Hb can be calculated.

The expected variance in log(Hb) informs the behaviour of the variance of the individual peaks and vice versa. Since the variance of a sum is the sum of the variances then:

$$E\{var[\log(\mathrm{Hb})]\} = E\left\{var\left[\log\left(\frac{E_{\mathrm{HMW}}}{E_{\mathrm{LMW}}}\right)\right]\right\}$$
$$= E\{var[\log(E_{\mathrm{HMW}}) - \log(E_{\mathrm{LMW}})]\}$$
$$= E\{var[\log(E_{\mathrm{HMW}})] + var[\log(E_{\mathrm{LMW}})]$$
$$-2\ \mathrm{covar}[\log(E_{\mathrm{HMW}}), \log(E_{\mathrm{LMW}})]\}$$

We assume that given a set of mass parameters (see [11] for the mass parameters considered in the STRmix™ model) the expected peak heights of two peaks of a heterozygous pair are independent and hence covar[log(E_{HMW}), log(E_{LMW})] = 0. We also assume that the two peaks of a heterozygous pair will have approximately the same variance so that var[log(E_{HMW})] = var[log(E_{LMW})] = var[log(E)] meaning that:

$$E\{var[\log(Hb)]\} = 2 \times E\{var[\log(E)]\}$$

i.e. the variance of log(Hb) is expected to be twice the variance of the log of the individual peaks making it up. The advantage of developing peak height variability as opposed to simply measuring the variability in Hb is that applying a Hb 'interpretation rule' to mixed DNA profiles quickly becomes problematic when shared peaks are present, whereas peak height variability can still readily be applied. There are different models for peak height variability modelling [9,11,24], but we demonstrate the idea using the model in line with [11] in which we model the observed and expected peak heights by:

$$\log\left(\frac{O}{E}\right) \sim N\left(0, \frac{c^2}{E}\right) \quad (3)$$

Using this model and the proof above, the bounds on Hb expected from a peak height variance constant of $c2$ can then be determined by:

$$\pm\alpha\sqrt{\frac{2c^2}{E}} \quad (4)$$

Where α is the critical value that determines the quantile of the bounds. We draw in Fig. 5 the 95% bounds ($\alpha = 1.96$) and adjust $c2$ value to 6.1 so that exactly 95% of the log(Hb) data is covered by these bounds. Note that APH in Fig. 5 is equivalent to E in Equation 3 for the purposes of graphing the bounds and log(Hb) together in Fig. 5.

The expected peak height variance for this data is $6.1/E$. The process we have gone through here is the reverse of how it is carried out in STRmix™, where peak height variability is determined directly from the data and then checked against Hb secondarily. Doing so provides a distribution for $c2$, for which the mean is 6.4 (data not shown).

2.5. Drop-in

Negative control samples are samples prepared with all the reagents and plasticware used for a case sample but with no actual case material added. These samples do occasionally show peaks, especially when using enhanced sensitivity methods. These peaks appear mainly as single peaks per control or less often two or three. Even more rarely a significant part of a profile appears [25]. It has been theorized that two mechanisms are in operation:

1. Tiny fragments of extraneous DNA, or
2. A full cell or a large part of a cell

is introduced into the PCR from the laboratory environment or consumables used [26].

The occasional peaks are referred to as drop-in. The more complete profiles are referred to as contamination. The frequency of detection of drop-in peaks increases as the sensitivity of the DNA testing method increases.

There are few published studies of the heights of drop-in alleles. Recently a model for drop-in peak heights was suggested [27]. This suggested that the height of a drop-in peak should follow a gamma distribution and the number of drop-ins modelled with a Poisson distribution, both of which are estimated from negative controls. The gamma distribution can have a very wide range of shapes depending on the parameters and hence can be used to explain many observed distributions.

The experimental design used to inform the drop-in model should be matched to the way that drop-in will be applied in casework. This should consider any AT used. What is required is data from a large number of positive observations, say 100, from negative controls, such as suggested in [27]. These should be analyzed to very low heights, say 10 RFU, regardless of what value is used for the AT in casework. The height of each peak should be recorded as well as the number of peaks per control. In order for the peaks to be considered drop-in they should not be reproducible on subsequent PCR of the same DNA extract.

For the semi-continuous models it is the rate of drop-in that is required. However the continuous models require both the rate and the peak height. The same experimental design can be used for either SEP but for the continuous models the data need to retain height information.

Table 3
Count of negative controls with 0–10 observed alleles.

Number of peaks per control	0	1	2	3	4	5	6	7	8	9	10
Count of controls with this number of peaks	325	101	24	8	2	3	2	0	1	1	0

Too few drop-in events were observed using GlobalFiler™ at 29 PCR cycles to demonstrate the gamma modelling, so we provide data generated using Life Technologies MiniFiler™ amplification kit performed at 30 PCR cycles to illustrate the approach taken. $N = 467$ negative control samples for the MiniFiler™ multiplex were read to 10 RFU using GeneMapper ID-X. In Table 3 are shown the observed counts of peaks per negative control.

The controls with high peak counts (to the right hand end of Table 3) would be considered contamination. Those to the left hand end would be described as drop-in. An arbitrary cut off between these two mechanisms, ε is often set.

For the data in Table 3, ε could be set at 2 or 3. This distinction is arbitrary but does have some scientific basis. We note that there is no "conservative" side to Pr(C) for all cases. It will vary for different cases. Hence it is not safe to err deliberately upwards or downwards.

In Fig. 6 is shown the observed number of peaks per profile (as a relative occurrence). This is overlayed with a Poisson distribution. The Poisson distribution is the expected distribution if the peaks were appearing independently as separate "drop-in" events. Also shown is a horizontal line that represents a constant contamination rate. The two lines cross at 3 peaks per negative control. This sets the definition of contamination at the point when the drop-in peaks no longer fit an independence model. For the demonstration set and setting $\varepsilon = 3$, those samples with 4 or more peaks were removed lowering N to 458 and giving $x = 173$ drop-in events.

If it is assumed that each drop-in observation is an independent event i.e. the probability of two drop-in events occurring is the product of each one occurring individually, then we need only a single drop-in probability, Pr(C). This can be calculated using data collated from the monitoring of negative controls samples tested within the laboratory by:

$$\Pr(C) = \frac{x}{N \times L}$$

Where x is the number of observed drop-in peaks above the level to be used for the AT (amended by S if positive samples are used), N is the number of profiles examined and L is the number of loci. For the data in Table 3 using AT = 10 RFU this is:

$$\Pr(C) = \frac{173}{458 \times 8} = 0.046$$

For the semi-continuous SEPs we would suggest that:

1. The counts per negative control should be recorded,
2. This should be plotted against a Poisson distribution and ε set,
3. x is then recalculated at the chosen AT and Pr(C) should be calculated.

For the continuous models it is necessary to consider the height of the drop-in peaks as well. In Fig. 7 is given the smoothed probability density of the observed distribution of drop-in peak heights overlaid with a gamma distribution fitted by MLE and scaled for the probability "missing" below 10 RFU.

2.6. Saturation

There is a well-recognised relationship between the amount of DNA that has been added to a PCR reaction (template) and the level of fluorescence of the peaks on the epg. The relationship has been shown to be approximately linear over a wide range of input DNA. Beyond this point the amount of fluorescence that is being produced by the excitation of the amplicon fluorophores is beyond the levels that can be detected by the charge coupled device (CCD) camera of an electrophoresis instrument. The CCD camera has become 'saturated' by input signal. This is a phenomenon more associated with the camera rather than the multiplex but we suggest that it should be validated for each multiplex and SEP.

To calculate the saturation point the relationship between stutter and parent peak heights can be used. As the amount of DNA added to a PCR reaction increases the heights of both the allele and its stutter will increase proportionally. The allelic peak, being much more intense than the stutter peak, will reach the saturation point before the stutter peak. From this point on the observed stutter ratio will diverge from its expected value.

Using the observed stutter peak height (O_{a-1}) and the stutter ratio for that allele and locus $\left(\mathrm{SR}_a^l\right)$, the expected parent peak height (E_a) can be calculated by:

$$E_a = \frac{O_{a-1}}{\mathrm{SR}_a^l}$$

And compared with the observed parent peak height (O_a). This analysis was carried out for the GlobalFiler™ dataset analysed on a 3130xl and can be seen in Fig. 8.

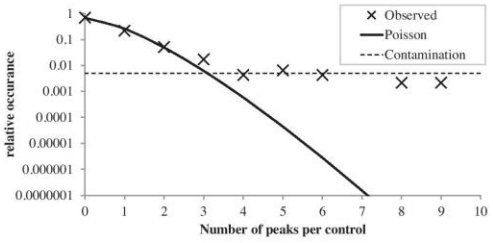


Fig. 6. A plot of the observed and expected number of peaks per profile (as relative occurrence), the Poisson distribution fitted to this data and the contamination rate.

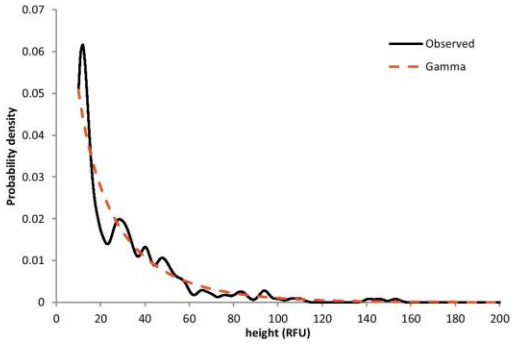


Fig. 7. Gamma distribution fitted to the observed MiniFiler™ (30 cycle, AT = 10 RFU) drop-in data.

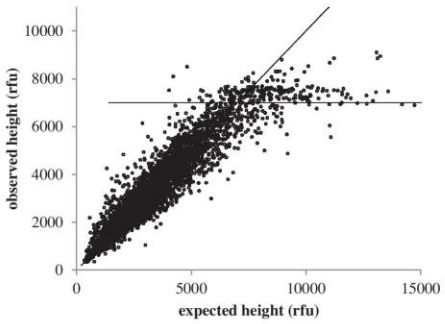


Fig. 8. Observed allele peak height compared to expected allele peak height based on its stutter peak RFU.

A deviation from the line of equality in Fig. 8 can be seen between the observed and expected peak heights around 7000 RFU. This means that a peak observed at 7000 RFU does not indicate the level of input DNA other than to suggest it is above some level that corresponds with the saturation point. Note that we desire a saturation level that sits immediately below the point at which observed peak heights plateau rather than an average value, as the point of the saturation level is that any peaks heights above it do not necessarily represent input DNA.

A further point of note from Fig. 8 is that the saturation does not appear to be a gradual process, rather one that occurs suddenly at a particular point. Therefore an appropriate model for fluorescence may be:

$$E_a = \begin{cases} f(M) & f(M) < s \\ s & \text{otherwise} \end{cases}$$

Where s is the saturation point and $f(M)$ is some function of mass parameters that predicts peak height.

2.7. Drop-out

The method used to model the probability of drop-out Pr(D) can vary significantly depending on the implementation of the drop-out model in an interpretation system. At writing there is quite a diversity of implementations. There are two main variants of drop-out modelling; that which treats drop-out as an extreme form of imbalance and that which treats drop-out events with their own model.

2.7.1. Drop-out considered as an extreme peak height imbalance

Drop-out can be considered an extreme form of imbalance where a peak has fallen below the AT [28]. This implies two things:

1. The model for peak balance should be able to be extended to handle drop-out events without the need for a separate drop-out model, and
2. The probability of drop-out will depend on the AT, cycle number, template, and amplification efficiency at that locus.

For a given AT, multiplex and cycle number, the primary determinant of drop-out is template. For a profile showing a typical degradation curve [29] we would expect a low probability of drop-out at the low molecular weight end and a higher one at the high molecular weight end in line with a reduction in intact template. In addition some loci amplify above or below the general amplification efficiency trend across all loci in the profile and this effect is not constant over time [30]. STRmix™ uses an extension of Eq. (3), which considered that the observed peak height for a dropped out allele has a uniform prior distribution between 0 and the AT. This is explained fully in [11] and we reproduce the operative formula below:

$$\Pr(O < \mathrm{AT}|E) = \int_{i=0}^{\mathrm{AT}} \Pr(O = i|E)\Pr(O = i)di$$

$$\text{where } \Pr(\mathrm{O} = \mathrm{i}|\mathrm{E}) \sim \mathrm{N}\left(0, \frac{c^2}{E}\right)$$

$$\text{and } \Pr(\mathrm{O} = i) \sim \mathrm{U}(0, \mathrm{AT})$$

Note that this method also works for instances when $E < \mathrm{AT}$.

2.7.2. Specific drop-out probability models

LRmix and LikeLTD do not use validation data to inform Pr(D), instead they use the number of observed alleles or maximum likelihood estimation respectively. Lab Retriever uses a variant of the logistic regression method of Tvedebrink et al. [31] without consideration of degradation. The Lab Retriever formula (LRF) is $\Pr(\mathrm{D}) = \frac{e^{\beta_0+\beta_1\overline{H}}}{1+e^{\beta_0+\beta_1\overline{H}}}$ (where $\overline{H}$ is a proxy for template and β_0 and β_1 are determined by MLE from training data) and differs from Tvedebrink's first version in that it does not use $\mathrm{Ln}(\overline{H})$. Tvedebrink went on to refine this version to incorporate the effects of degradation [30–32]. The Forensic Statistical Tool developed by the Office of the Chief Medical Examiner of the City of New York uses a system based on quantification estimate [8]. Recently investigations into drop-out modelling have included DNA amounts, PCR cycle number and capillary electrophoresis injection time as dependent variables in a drop-out probability model [33].

2.7.3. Comparing models to observations

It has been suggested from theoretical studies [34] that variability is introduced early in the process when an aliquot is taken from the extract. In the extract the cells are ruptured and the aliquot may take a varying number of template DNA molecules. The observed fraction of dropped alleles is expected to be the result of these processes and stochastic variation. We will term the observed fraction of dropped alleles as the drop-out frequency for that profile. This can be observed.

There are a number of methods that can be employed to observe drop-out, which rely on the height of surviving peaks in a heterozygous pair or in a profile. We demonstrate examination of these models using nine low level GlobalFiler™ profiles. Quantification estimates for these profiles indicated the total input template ranged from 50 to 210 pg. The profiles showed only modest degradation slopes. The profiles were read down to 10 RFU which allowed the examination of Pr(D) for thresholds at 10, 30 and 50 RFU. Since Lab Retriever uses the average peak height from the profile to create the template proxy we split the nine profiles into a set of four and a set of five. One set was used to train the logistic model which was then trialled on the other set. The training and test sets were then reversed and the process repeated. The results are shown in Fig. 9 below.

In the same study where the LRF was trialled we also trialled, $\Pr(\mathrm{D}) = \frac{e^{\beta_0+\beta_1 \ln Q}}{1+e^{\beta_0+\beta_1 \ln Q}}$ and $\Pr(\mathrm{D}) = \frac{e^{\beta_0+\beta_1 Q}}{1+e^{\beta_0+\beta_1 Q}}$ where Q was the quantification estimate. These two variants using the quantification estimate were both highly ineffective at modelling Pr(D).

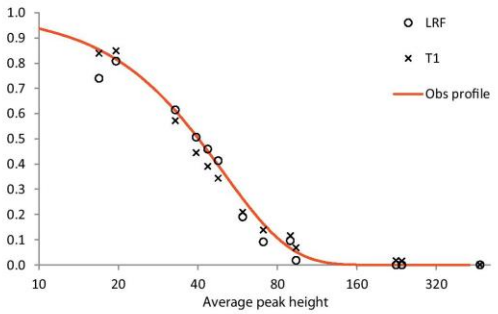


Fig. 9. A plot of expected drop-out probabilities using the Lab Retriever formula (LRF) and Tvedebrink's 1st model (T1) formulae for thirteen low level GlobalFiler™ mixtures at a threshold 30 RFU vs the observed drop-out probabilities, plotted against APH. The observed line is hand drawn to the data (not shown).

When examining the STRmix™ model we cannot compare with the observed drop-out in a profile as we did for LRF and T1 in Fig. 9. This is because STRmix™ has a drop-out probability for each allele. In addition STRmix™ does not have one value even for each allele, rather it integrates the probability of the profile across varying Pr (D). In Fig. 10 we give an approximation to the mean Pr(D) for STRmix™ vs the moving average of the observed drop-out at the template proxies (this can be thought of as the expected peak height of the allele).

To our eye both drop-out approximations shown in Fig. 9 and the STRmix™ approximation shown in Fig. 10 appears to fit expected to observed data well.

2.7.4. Using drop-out models to develop thresholds

In many instances drop-out probability models will be required to develop thresholds to determine when the probability of drop-out is sufficiently improbable that the expert is willing to round it down to zero. This threshold is often termed the homozygous threshold, drop-out threshold or stochastic threshold and is used to determine when a single peak at a locus represents (with the desired confidence) a homozygous genotype. One method for setting a drop-out threshold was given in [35]. Using the STRmix™ method the value of E can be chosen to equate to that point at which the cumulative normal distribution equals the desired Pr(D) value.

Other methodologies may require separate modelling of drop-out events in order to develop a homozygous threshold as the

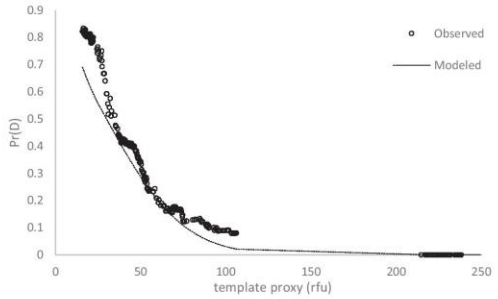


Fig. 10. A plot of expected and observed drop-out probabilities (at 30 RFU) vs a template proxy using an approximation to the STRmix™ algorithm for thirteen low level GlobalFiler™ mixtures.

models that incorporate Pr(D) into their probabilistic systems are not manipulable in the same manner.

3. Training for continuous interpretation systems

Many authorizing bodies and courts are not only interested in the validity of the software but whether they can be confident that it is competently applied. To this end the provision of training and competence testing is vital. Admissibility hearings tend to focus on the reliability and acceptance of the method, and by extension the individual software programs used for the semi and fully continuous probabilistic software options. However, it is imperative to note that to have a successful SEP in use for forensic casework requires not only a validated software tool, but a fully trained and competent expert that uses the chosen software.

When discussing training for expert testimony we could start with rule 702 of the Federal Rules of Evidence governing admissibility of expert testimony in the US. This rule allows expert opinion evidence if:

A witness who is qualified as an expert by knowledge, skill, experience, training, or education may testify in the form of an opinion or otherwise if:

a) the expert's scientific, technical, or other specialized knowledge will help the trier of fact to understand the evidence or to determine a fact in issue;
b) the testimony is based on sufficient facts or data;
c) the testimony is the product of reliable principles and methods; and
d) the expert has reliably applied the principles and methods to the facts of the case.

We draw from this rule that it is the witness, not the software, that is qualified by the court as an expert.

The right to confront adverse witnesses is ancient. It appears in the Acts of the Apostles 25:16, when Roman governor Porcius Festus states when discussing the proper treatment of Paul: "It is not the manner of the Romans to deliver any man up to die before the accused has met his accusers face-to-face, and has been given a chance to defend himself against the charges." It is also cited in Shakespeare's *Richard II:* "Then call them to our presence; face to face, And frowning brow to brow, ourselves will hear The accuser and the accused freely speak."

The European Court of Human Rights, Article 6(3), provides that 'everyone charged with a criminal offence' has the right to 'examine or have examined witnesses against him'. This basically means that the accused, or his lawyer, should have a chance to put questions to adverse witnesses. The Sixth Amendment to the Constitution of the United States of America provides that a person accused of a crime has the right to confront a witness against him or her in a criminal action. This includes the right to be present at the trial as well as the right to cross-examine the prosecution's witnesses. It is therefore essential that the witness can represent the evidence and meet the needs of cross examination.

It is accepted that no analyst is required to understand the mathematics and computer program to the extent that they could recreate the system, except the developers themselves. However it is an expectation that analysts at least understand the workings of any system they use to be able to understand and explain the results.

This idea was demonstrated during the trial in *R v Noll* (*R v Noll* (1999) 3 VR 704) when the witness acknowledged that although his evidence was based on accepted scientific theory, he himself could not describe that theory. During an appeal it was submitted that this meant the witness was incapable of giving the DNA evidence and should have been excluded. The court found that

Table 4
Evidence of acceptance for some principles underlying the STRmix™ software.

Principle	Evidence of acceptance
Monte Carlo Markov Chain	This is very standard statistical method employed in many areas of science. Searching the term "Monte Carlo Markov Chain" in Scopus returns more than 22,000 records. Scopus is an online bibliographic database containing abstracts and citations for 20,000 peer-reviewed academic journals' articles.
Stutter and peak heights and the variance about them can be predicted from empirical models	Studies of stutter and allele peak heights and now quite numerous and have appeared in the peer reviewed literature [14,16,38].
The probability of a multilocus genotype can be estimated from allele probabilities and the coancestry coefficient.	For STRmix™ the model follows Balding and Nichols [39]. This model is based on published literature and appears as NRC II recommendation 4.2. It is the most conservative of the methods in common forensic use [40].

although the witness was unable to explain the technical aspects of the theory, he was entitled to rely on other expert opinion. Addressing this issue specifically Ormiston J explains:

> "Professional people in the guise of experts can no longer be polymaths; they must, in this modern era, rely on others to provide much of their acquired expertise. Their particular talent is that they know where to go to acquire that knowledge in a reliable form."

As yet there is no simple definition of the level of technical knowledge required for an analyst to be considered competent to use an interpretational system. We outline our own opinion[2] here considering STRmix™ on what we consider essential to effectively represent the evidence.

- Likelihood ratios
- Choosing propositions suitable for the case and the hierarchy of propositions
- MCMC and the Metropolis-Hastings algorithm
- Peak and stutter height models
- Balding and Nichols formulae
- Limits and uncertainties of an LR produced with STRmix™
- Diagnosing poor performance.

STRmix™ is *LR* based and we would suggest that understanding of likelihood ratios is an essential first step. The switch to likelihood ratios from, say, exclusion probabilities may be one of the more challenging aspects of the training. Training in this should cover how to tailor propositions to the case in question and how to recognise when there is a shift in the position in the hierarchy of propositions. There should be training on the prosecutor's fallacy [36] and methods to avoid making this error.

Training should be undertaken on written report wording and on verbal testimony. Moot court is appropriate and should be robust. In verbal testimony a balance must be maintained between understanding and thoroughness. Hence an ability to use lay terms is essential. The witness should also be tested on the ability to recognise his own limitations and to explain this at the appropriate time in testimony.

Because the LR is based on competing propositions, the wording of the findings must clearly state what propositions were used in performing the LR calculation. Again, because no two cases in forensics are exactly the same, the training should focus on how to clearly convey the findings rather than some pre-set determination of which set of propositions would be a lab wide "default LR" and boiler plate type reporting statements.

LRs are often given at the sub-source level but court questioning often progresses seamlessly to the activity level. The witness should be able to recognize a shift in level of the hierarchy of propositions and to adjust testimony appropriately. She should also be able to explain the transition in lay terms.

For the purposes of representing STRmix™ assisted LRs the witness should be able to explain MCMC and the Metropolis–Hastings algorithm in lay terms. Analogies to hill mapping may be appropriate. Training should cover Monte Carlo variability and its implications for testimony. An ability to explain the underlying peak and stutter heights models is essential but this is usually well within the existing knowledge of many experienced analysts.

Continuous approaches all have in common the ability to use the entire profile, resulting in very few times where a questioned sample is inherently not suitable for a statistical calculation. This has required a re-think of the definitions of inclusion, exclusion and inconclusive. The ability of the software to produce a number does not absolve the expert from the duty of recognizing when the software should be used at all nor whether it has run optimally. It is therefore essential that training on the limitations of the method is given and received.

4. General acceptance of continuous interpretation systems

We consider here the general acceptance of the subject of probabilistic genotyping. Most jurisdictions have rules governing the introduction of new scientific concepts. In the US this is governed variously by the Frye and Daubert standards. In this section we discuss only one part of that standard, often termed general acceptance. The Frye standard [37] arises from the case *Frye v. United States*, 293 F. 1013 (D.C. Cir. 1923) in which the court gave the opinion:

> Just when a scientific principle or discovery crosses the line between the experimental and demonstrable stages is difficult to define. Somewhere in this twilight zone the evidential force of the principle must be recognized, and while the courts will go a long way in admitting experimental testimony deduced from a well-recognized scientific principle or discovery, the thing from which the deduction is made *must be sufficiently established to have gained general acceptance* in the particular field in which it belongs.

This emphasizes that the deduction must proceed from a well-recognized scientific principle or discovery. Moving to software this would appear to mean that the software must implement accepted scientific principles. We would not read this as meaning that the software itself must be in prevalent use but that the principles upon which it is based must be generally accepted. This is thoroughly sensible. Obviously when any software first appears it will be in limited use but it may be very soundly programmed from well accepted principles. The court clearly envisages that the standard is that the principles are sound, not some sort of vote about how often the software is used. Having said that, it is clear that training must explain the correct use of the software in a manner that is appropriate to the expert's role in using the software. If an end user must participate in an admissibility

[2] Much of the following list and commentary is influenced by experience at the Defense Forensic Science Center (DFSC) who have recently implemented a probabilistic approach. The views expressed here those of the authors and not official policy of the Departments of Defense or Army.

hearing, this may require a greater depth of training than is needed for testimony on the use and results.

Since this is a paper on validation this suggests that any developer should outline the principles upon which the software is based and ensure that these meet the standard.

We outline these for the STRmix™ software as an example in Table 4.

Agreement in science proceeds by the peer reviewed literature. We have surveyed the literature using the Scopus online search tool. We searched for the key words "forensic and DNA and interpretation" 2012-present and obtained 150 references. These were scored as pro the use of probabilistic genotyping; anti or irrelevant. We obtained 39 pro; 1 anti [41]; and 110 that were not relevant. A key phrase from the one scored as anti was: "The variance of heterozygote balance was more expanded in two-person mixtures than in one person samples. Therefore; it is not suitable to use allelic peak heights/areas for estimating the genotypes of the contributors such as the quantitative analysis."

Of note, the ISFG 2012 Guidelines recommend probabilistic methods. In the US, both the Scientific Working Group on DNA Analysis Methods (SWGDAM) and the Organization of Scientific Area Committees (OSAC) DNA Analysis 2 sub-committee are working on guidelines for assessing probabilistic genotyping software tools. In addition, a considerable number of modern probabilistic genotyping software programs are developed or being developed by researchers or academics often with very strong mathematical or statistical backgrounds [6,8,9,11,12,24,42–45].

5. Conclusion

Forensic laboratories are moving from binary, threshold based systems towards semi-continuous and fully continuous alternatives. A necessary prerequisite of this transition is the development of models that describe DNA profiles behaviours. Currently a number of models exist and are implemented in different ways that suit the particular interpretation system being used. Perhaps this will always be the case, or perhaps there will be convergence to a common system of models that prove superior to all others. Regardless of which direction the future takes, work is required to develop, refine and test potential models. We emphasise the necessity of a fit between the validation of a multiplex and the intended interpretation methodology.

We provide in this work some approaches to modelling for a number of aspects of DNA profile behaviour. Some of these ideas are already in use in active casework.

Some modelling given here still provides thresholds (e.g. analytical threshold), which would ideally (but not yet practically) be completely removed from continuous DNA profile interpretations. Models continually improve and often the improvement of one aspect of a model will reveal a path to further improvements or refinements elsewhere. These improvements should not be viewed as evidence that the previously used model is unreliable. The nature of science is one of continual improvement, with each model being tested and, if found to perform 'better' in some desirable way, implemented into active use. At this point the new model becomes the best practise available at the time. Any argument which seeks to nullify all previous work purely because better models are developed is an argument against ever providing any results and only works to generate disincentives for progress and improvement.

Acknowledgements

This work was supported in part by grant 2011-DN-BX-K541 from the US National Institute of Justice. Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the U.S. Department of Justice or the Departments of Defense and Army. Names of commercial manufacturers or products included are incidental only, and inclusion does not imply endorsement by the Departments of Defense.

Appendix 1. An approach to the calculation of the AT

1. An approach to the calculation of the AT

GeneMapper™ ID-X baselines were measured by noting all fluorescence (i.e. the RFU value) every 5 bp between 70 bp and 460 bp in all dye lanes for 5 GlobalFiler™ profiles where template values ranged from 0 pg (negative controls) to those produced using 500 pg of target DNA. This was carried out on two different 3130xl capillary electrophoresis instruments. Where a reading point corresponded to a peak (whether artefactual or allelic) then this point was considered missing data. Artefactual peaks removed at analysis included back and forward stutter peaks, pull-up and those within known dye blob positions. This resulted in 4063 measured values.

The average and standard deviation (sd) of these data were calculated for each template, dye colour, and for each of the two capillary electrophoresis machines. These averages and sd were regressed against dye color, template, and machine using multiple regression in EXCEL. The results are shown in Table A1.

The regression analysis suggests small but significant effects of color and machine on both the mean and standard deviation (sd) of noise. There also appears to be an effect of template on sd. This regression assumes normality of baseline fluorescence and as we show later a gamma distribution appears to explain the data equally well. We therefore consider how the choice of distribution may affect the size of the calculated AT thresholds.

In Table A2 are given the predicted values of the mean +10sd and 0.99999 quantile of the gamma distribution for the two different machines, for the different dye colors, at 10 and 500 pg template values. The value of the mean +10sd has been suggested as one way to calculate the AT [46]. The data in Table 2 under the gamma distribution 0.99999 quantile have been developed by fitting a gamma distribution to the data using maximum likelihood estimation, (MLE). The parameters of the gamma were allowed to vary with template.

Looking at the values of the mean +10sd columns in Table A2 it is possible to pick one AT value to encompass both machines and all colors and most templates. This could be 30 RFU or even lower values could be considered. No actual data in our demonstration set exceed this value. The maximum observed was 14 RFU.

Modelling baseline in this manner assumes normality of the noise, and this has been suggested as approximately true [47]. In

Table A1
The results of the regression analysis.

	Mean		Standard deviation	
	Coefficients	*p*-Value	Coefficients	*p*-Value
Intercept	1.36	8.5E-16	1.29	8.7E-22
Blue	0.35	1.9E-02	0.08	4.4E-01
Green	1.39	4.2E-13	0.47	1.3E-05
Orange	0.73	6.3E-06	0.11	2.7E-01
Purple	0.96	2.2E-08	0.37	3.7E-04
Red	0.68	2.0E-05	0.12	2.3E-01
Yellow was used as the benchmark hence its coefficient is 0.00				
Machine	0.53	5.0E-08	0.25	4.7E-05
Template (pg)	0.0002	3.9E-01	0.0011	9.1E-10

Table A2
Mean + 10SD and the Gamma distribution 0.99999 quantile for two different machines, for the different colours, at 10 and 500 pg template.

Machine	Colour	10 pg		500 pg	
		Mean + 10sd	Gamma 0.99999 quantile	Mean + 10sd	Gamma 0.99999 quantile
1	Blue	19	17	24	17
	Green	24	19	29	26
	Orange	19	15	25	15
	Purple	22	16	27	20
	Red	19	20	25	22
	Yellow	18	19	23	19
2	Blue	16	15	21	15
	Green	21	20	26	19
	Orange	16	13	22	15
	Purple	19	15	24	16
	Red	16	15	22	19
	Yellow	14	14	20	15

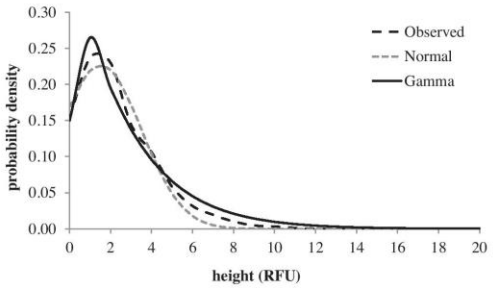


Fig. A1. Observed peak heights of noise peaks on two 3130xl machines using GeneMapper™ ID-X software with a normal and a gamma distribution overlaid.

Fig. A1 we give a plot of the observed data across all colors, template amounts, and machines. The observed line in Fig. A1 was constructed by drawing a line through the observed proportions of baseline fluorescence that occurred in each 1 rfu bracket. Also in this graph is the best fit normal distribution. Our data show a very slight positive skew as evidenced by the small departures of the observed from the best fit normal distribution. We have investigated whether a gamma distribution might fit better. The best fit gamma is also shown in Fig. A1 and subjectively looks to be a similar fit to the normal except that it has the advantage of being bounded by 0 and hence has no density at negative values. We are likely to be interested in the positive tail. On these data the gamma slightly overestimates the density at the positive tail whereas the normal slightly underestimates it.

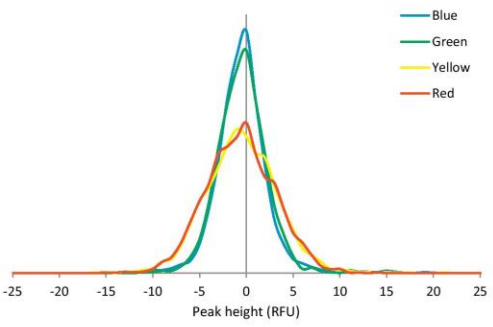


Fig. A2. Observed RFU level of baseline noise data points using Osiris™ software on one 3130xl machine in five positive Identifiler Plus samples. The colours are the Identifiler Plus dye colours.

The OSIRIS™ [48] software uses a different baselining algorithm. Osiris™ displays rfu on a scale that can be negative as it sets an rfu level of zero as the average of the baseline. In Fig. A2 we give the distribution of peak heights of noise peaks in five positive samples. In these five samples there were 25 regions above 10 RFU with the largest being 21 RFU. At least some of these had acceptable morphology.

This demonstrates the factors that we have identified need examination during validation. Specifically, the baseline noise should be examined for every color, each machine in use, and over template ranges. The differences observed in this dataset were in many cases statistically significant but small. It would be reasonable when setting an AT to either have:

1. One AT for all machines, colors and templates, or
2. Different values for each machine and color, plausibly set near the optimal template value.

When setting an AT, the impact on time spent at CE data analysis may well be a necessary practical consideration. For very low ATs it is likely additional time will be spent manually removing bad morphology peaks. As stated in the main text, electronic noise is only one part of a larger problem.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.fsigen.2015.09.011.

References

[1] D.R. Hares, Expanding the CODIS core loci in the United States, Forensic Sci. Int. Genet. 6 (2012) e52–e54.
[2] D.R. Hares, Addendum to expanding the CODIS core loci in the United States, Forensic Sci. Int. Genet. 6 (2012) .
[3] L.A. Welch, P. Gill, C. Phillips, R. Ansell, N. Morling, W. Parson, et al., European Network of Forensic Science Institutes (ENFSI): evaluation of new commercial STR multiplexes that include the European Standard Set (ESS) of markers, Forensic Sci. Int. Genet. 6 (2012) 819–826.
[4] T. Schellberg, N. Oldroyd, L.L. Schade, Maximizing the power of forensic DNA databases with next generation STR technology, Forensic Magaz. (2012) http://www.forensicmag.com/articles/2012/10/maximizing-power-forensic-dna-databases-next-generation-str-technology.
[5] Scientific_Working_Group_on_DNA_Analysis_Methods_(SWGDAM). SWGDAM Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Testing Laboratories. 2010.
[6] P. Gill, H. Haned, A new methodological framework to interpret complex DNA profiles using likelihood ratios, Forensic Sci. Inte. Genet. 7 (2013) 251–263.

[7] C.D. Steele, M. Greenhalgh, D.J. Balding, Verifying likelihoods for low template DNA profiles using multiple replicates, Forensic Sci. Int. Genet. 13 (2014) 82–89.
[8] A.A. Mitchell, J. Tamariz, K. O'Connell, N. Ducasse, Z. Budimlija, M. Prinz, et al., Validation of a DNA mixture statistics tool incorporating allelic drop-out and drop-in, Forensic Sci. Int.: Genet. 6 (2012) 749–761.
[9] R. Puch-Solis, T. Clayton, Evidential evaluation of DNA profiles using a discrete statistical model implemented in the DNA LiRa software, Forensic Sci. Int. Genet. 11 (2014) 220–228.
[10] J.-A. Bright, I.W. Evett, D. Taylor, J.M. Curran, J. Buckleton, A series of recommended tests when validating probabilistic DNA profile interpretation software, Forensic Sci. Int. Genet. 14 (2015) 125–131.
[11] D. Taylor, J.-A. Bright, J. Buckleton, The interpretation of single source and mixed DNA profiles, Forensic Sci. Int. Genet. 7 (2013) 516–528.
[12] M.W. Perlin, M.M. Legler, C.E. Spencer, J.L. Smith, W.P. Allan, J.L. Belrose, et al., Validating TrueAllele® DNA mixture interpretation, J. Forensic Sci. 56 (2011) 1430–1447.
[13] C. Rakay, J. Bregu, C. Grgicak, Maximizing allele detection: Effects of analytical threshold and DNA levels on rates of allele and drop-out, Forensic Sci. Int. Genet. 6 (2012) 723–728.
[14] J.-A. Bright, J.M. Curran, J.S. Buckleton, Investigation into the performance of different models for predicting stutter, Forensic Sci. Int. Genet. 7 (2013) 422–427.
[15] J.-A. Bright, K.E. Stevenson, M.D. Coble, C.R. Hill, J.M. Curran, J.S. Buckleton, Characterising the STR locus D6S1043 and examination of its effect on stutter rates, Forensic Sci. Int. Genet. 8 (2014) 20–23.
[16] J.-A. Bright, D. Taylor, J.M. Curran, J.S. Buckleton, Developing allelic and stutter peak height models for a continuous method of DNA interpretation, Forensic Sci. Int. Genet. 7 (2013) 296–304.
[17] C. Brookes, J.-A. Bright, S. Harbison, J. Buckleton, Characterising stutter in forensic STR multiplexes, Forensic Sci. Int. Genet. 6 (2012) 58–63.
[18] J.-A. Bright, J.S. Buckleton, D. Taylor, M.A.C.S.S. Fernando, J.M. Curran, Modelling forward stutter: towards increased objectivity in forensic DNA interpretation, Electrophoresis 35 (2014) 3152–3157.
[19] R Core Team, R: A language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2013.
[20] H. Kelly, J.-A. Bright, J.M. Curran, J. Buckleton, Modelling heterozygote balance in forensic DNA profiles, Forensic Sci. Int. Genet. 6 (2012) 729–734.
[21] J.-A. Bright, J. Turkington, J. Buckleton, Examination of the variability in mixed DNA profile parameters for the Identifiler(TM) multiplex, Forensic Sci. Int. Genet. 4 (2009) 111–114.
[22] J.-A. Bright, K. McManus, S. Harbison, P. Gill, J. Buckleton, A comparison of stochastic variation in mixed and unmixed casework and synthetic samples, Forensic Sci. Int. Genet. 6 (2012) 180–184.
[23] T. Tvedebrink, H.S. Mogensen, M.C. Stene, N. Morling, Performance of two 17 locus forensic identification STR kits—Applied Biosystems's AmpF/STR® NGMSElect™ and Promega's PowerPlex® ESI17 kits, Forensic Sci. Int. Genet. 6 (2012) 523–531.
[24] R.G. Cowell, S.L. Lauritzen, J. Mortera, Probabilistic modelling for DNA mixture analysis, Forensic Sci. Int. Gene. Suppl. Ser. 1 (2008) 640–642.
[25] P. Gill, J.P. Whitaker, C. Flaxman, N. Brown, J.S. Buckleton, An investigation of the rigor of interpretation rules for STR's derived from less that 100 pg of DNA, Forensic Sci. Int. 112 (2000) 17–40.
[26] J.S. Buckleton, C.M. Triggs, S.J. Walsh, Forensic DNA Evidence Interpretation, CRC Press, Boca Raton, Florida, 2005.
[27] R. Puch-Solis, A dropin peak height model, Forensic Sci. Int. Genet. 11 (2014) 80–84.
[28] P. Gill, L. Gusmão, H. Haned, W.R. Mayr, N. Morling, W. Parson, et al., DNA commission of the International Society of Forensic Genetics: recommendations on the evaluation of STR typing results that may include drop-out and/or drop-in using probabilistic methods, Forensic Sci. Int. Genet. 6 (2012) 679–688.
[29] J.-A. Bright, D.J.M. Taylor, CB, J.S. uckleton, Degradation of forensic DNA profiles, Aust. J. Forensic Sci. 45 (2013) 445–449.
[30] J. Buckleton, H. Kelly, J.-A. Bright, D. Taylor, T. Tvedebrink, J.M. Curran, Utilising allelic dropout probabilities estimated by logistic regression in casework, Forensic Sci. Int. Genet. 9 (2014) 9–11.
[31] T. Tvedebrink, P.S. Eriksen, M. Asplund, H.S. Mogensen, N. Morling, Allelic drop-out probabilities estimated by logistic regression—further considerations and practical implementation, Forensic Sci. Int. Genet. 6 (2012) 263–267.
[32] T. Tvedebrink, P.S. Eriksen, H.S. Mogensen, N. Morling, Statistical model for degraded DNA samples and adjusted probabilities for allelic drop-out, Forensic Sci. Int. Genet. 6 (2012) 97–101.
[33] R. Hedell, C. Dufva, R. Ansell, P. Mostad, J. Hedman, Enhanced low-template DNA analysis conditions and investigation of allele dropout patters, Forensic Sci. Int. Genet. 14 (2015) 61–75.
[34] P. Gill, J. Curran, K. Elliot, A graphical simulation model of the entire DNA process associated with the analysis of short tandem repeat loci, Nucleic Acids Res. 33 (2005) 632–643.
[35] R. Puch-Solis, A.J. Kirkham, P. Gill, J. Read, S. Watson, D. Drew, Practical determination of the low template DNA threshold, Forensic Sci. Int. Genet. 5 (2011) 422–427.
[36] W.C. Thompson, E.L. Schumann, Interpretation of statistical evidence in criminal trials - The prosecutors fallacy and the defence attorneys fallacy, Law Human Behav. 11 (1987) 167–187.
[37] Frye v The United States of America, 54 AppDC 46, 293Fed 1013 (1923). 1923.
[38] J.-A. Bright, J.M. Curran, Investigation into stutter ratio variability between different laboratories, Forensic Sci. Int. Genet. 13 (2014) 79–81.
[39] D.J. Balding, R.A. Nichols, DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands, Forensic Sci. Int. 64 (1994) 125–140.
[40] J.M. Curran, J.S. Buckleton, C.M. Triggs, What is the magnitude of the subpopulation effect? Forensic Sci. Int. 135 (2003) 1–8.
[41] S. Manabe, Y. Mori, C. Kawai, M. Ozeki, K. Tamaki, Mixture interpretation: experimental and simulated reevaluation of qualitative analysis, Leg. Med. 15 (2013) 66–71.
[42] R. Puch-Solis, L. Rodgers, A. Mazumder, S. Pope, I. Evett, J. Curran, et al., Evaluating forensic DNA profiles using peak heights, allowing for multiple donors, allelic dropout and stutters, Forensic Sci. Int. Genet. 7 (2013) 555–563.
[43] R.G. Cowell, S.L. Lauritzen, J. Mortera, Probabilistic expert systems for handling artifacts in complex DNA mixtures, Forensic Sci. Int. Genet. 5 (2011) 202–209.
[44] L. Prieto, H. Haned, A. Mosquera, M. Crespillo, M. Alemañ, M. Aler, et al., Euroforgen-NoE collaborative exercise on LRmix to demonstrate standardization of the interpretation of complex DNA profiles, Forensic Sci. Int. Genet. 9 (2014) 47–54.
[45] K. Lohmueller, N. Rudin, Calculating the weight of evidence in low-template forensic DNA casework, J. Forensic Sci. 58 (2013) 234–259.
[46] J.R. Gilder, K. Inman, W. Shields, D.E. Krane, Magnitude-dependent variation in peak height balance at heterozygous STR loci, Int. J. Legal Med. 125 (2011) 87–94.
[47] J. Bregu, Investigation of Baseline Noise: Establishing an RFU Threshold for Forensic DNA Analysis, Boston University, 2009.
[48] R.M. Goor, L.F. Neall, D. Hoffman, S.T. Sherry, A Mathematical Approach to the Analysis of Multiplex DNA Profiles, Bull. Math. Biol. 73 (2010) 1909–1931.

2.5 – Clarification

Regarding the dropout model:

Observed peak heights are indicated as *O*. When a peak has not been observed above a threshold (called an analytical threshold and designated as AT) then this is called a 'dropout'. We consider that *O* in this instance can take any value up to AT, $\left[O \mid O < AT\right] \sim U\left(0, AT\right)$. *O* is a random variable that has a log-normal distribution:

$$O \sim LN\left(\xi \log_{10}\left(E\right), \xi^2 \frac{c^2}{E}\right)$$

Where $\xi = \ln(10)$ is used to transform between logs in base 10 and base *e*. The probability of dropout is therefore obtained by:

$$\Pr(O < AT \mid E) = \int_0^{AT} p\left(o \mid E\right) do$$

In practice this can be achieved by summation over steps in *O* of 1rfu:

$$\Pr(O < AT \mid E) = \sum_{o=0}^{AT-1} p\left(o + 0.5 \mid E\right) \Pr\left(o < O < o + 1\right)$$

Note that we apply a transformation from modelling *O* directly to modelling $\log_{10}\left(\frac{O}{E}\right)$ by:

$$\log_{10}\left(\frac{O}{E}\right) \sim N\left(0, \frac{c^2}{E}\right)$$

The integral required to calculate the probability of dropout using the transformed variable, to provide equivalent dropout probability to the untransformed variable, is:

$$\Pr(O < AT \mid E) = \int_0^{AT} LN\left(o \mid \xi \log_{10}\left(E\right), \xi^2 \frac{c^2}{E}\right) do = \int_0^{\log_{10}\left(\frac{AT}{E}\right)} N\left(o \mid 0, \frac{c^2}{E}\right) do$$

2.6: Putting all the models together in a Bayesian framework for profile deconvolution using Markov Chain Monte Carlo

Manuscript: The interpretation of single source and mixed DNA profiles. D Taylor, JA Bright, J Buckleton. (2013) Forensic Science International: Genetics 7 (5), 516-528 – *Cited 75 times*

Statement of novelty: This paper describes the combination of modelling elements from the previous papers in this chapter. The method described in based on a Markov Chain Monte Carlo and the paper explains how the use of this method ultimately translates to a likelihood ratio (being the standard evidential weight used in forensic biology).

My contribution: I was the main author of the paper and equally responsible for theory and mathematics that the work is based on. I was the sole individual who programmed the software for simulations and analyses carried out.

Research Design / Data Collection / Writing and Editing = 45% / 60% / 55%

Additional comments:

Forensic Science International: Genetics 7 (2013) 516–528

Contents lists available at SciVerse ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

Forensic population genetics – original research

The interpretation of single source and mixed DNA profiles

Duncan Taylor [a], Jo-Anne Bright [b], John Buckleton [b,*]

[a] Forensic Science South Australia, 21 Divett Place, Adelaide, SA 5000, Australia
[b] ESR Ltd, Private Bag 92021, Auckland 1142, New Zealand

ARTICLE INFO

Article history:

Keywords:
Forensic DNA interpretation
Mixtures
Low template DNA
Continuous models

ABSTRACT

A method for interpreting autosomal mixed DNA profiles based on continuous modelling of peak heights is described. MCMC is applied with a model for allelic and stutter heights to produce a probability for the data given a specified genotype combination. The theory extends to handle any number of contributors and replicates, although practical implementation limits analyses to four contributors.

The probability of the peak data given a genotype combination has proven to be a highly intuitive probability that may be assessed subjectively by experienced caseworkers. Whilst caseworkers will not assess the probabilities per se, they can broadly judge genotypes that fit the observed data well, and those that fit relatively less well. These probabilities are used when calculating a subsequent likelihood ratio.

The method has been trialled on a number of mixed DNA profiles constructed from known contributors. The results have been assessed against a binary approach and also compared with the subjective judgement of an analyst.

1. Introduction

1.1. The likelihood ratio

A summary of notation used within this paper is provided in Appendix A. In forensic DNA work when a probative DNA profile is obtained from an evidentiary item it is normal to provide some assessment of the weight of evidence. This is commonly achieved by the presentation of a numerical statistic which is sometimes accompanied by an explanation in words. Two calculations are frequently used by the forensic community. These are the likelihood ratio (*LR*) and the combined probability of inclusion (*CPI*), also referred to as Random Man Not Excluded or *RMNE*. More recently a variant of the exclusion probability, termed the random match probability, has been adopted by many laboratories. We adopt an *LR* approach here.

Traditionally, the DNA profile is in the form of an electropherogram (EPG). The EPG is usually pre-processed using a set of heuristics [1], such as limit of detection or analysis thresholds.

As a general principle ignoring information that can be properly evaluated tends to weaken the evidence for a true hypothesis and will more often include a false hypothesis. Information, even relevant information, that cannot be properly evaluated may not have these effects and may simply add a random element to the analysis. Relevant information that can be effectively evaluated should not be ignored. In the *LR* framework including relevant and properly evaluated information tends to increase the *LR*s if H_1 is true and decrease the *LR*s if H_2 is true.

Certainly the first part of this principle has been elegantly reinforced by Perlin et al. [2] in their experiments with the continuous approach that makes efficient use of the information. In Section 6 we present a trial that supports Perlin et al.'s conclusion but also explores the second part of the prediction of this principle, that the use of more information should produce lower *LR*s when H_2 is true.

1.2. Lack of forward movement in DNA interpretation

Alternatives to the binary model were experimented with in the late 1990s [3–6]. Two options were tabled:

1. The fully continuous model, and
2. A model that is partially continuous based on allowing a probability for dropout and drop-in (hereafter the "drop model") [7].

These methods make assumptions about the variability of peak height, and the ratio of stutter peak height to allelic peak height (usually termed stutter ratio and given the symbol π here since *SR* may look like a multiplication of *S* and *R*).

* Corresponding author. Tel.: +64 9 8153 904; fax: +64 9 8496 046.
E-mail address: john.buckleton@esr.cri.nz (J. Buckleton).

1872-4973/$ – see front matter © 2013 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.fsigen.2013.05.011

Neither of these models saw any large-scale deployment throughout the early 2000s. This has come in for justifiable criticism [8–10]. However there has been a new and highly welcome forward movement in the late 2000s driven by the creation of continuous software (TrueAllele®), developed by Perlin et al. [2], and efforts from Balding and Buckleton [11], Haned and Gill [12] and Rudin and Lohmueller [13]. This has seen these techniques pass through the court process in both the UK and the US. TrueAllele® results have been reported in over 75 criminal cases. However progress is still partial [14] with only a few laboratories worldwide implementing or investigating fully continuous methods.

The barriers that had hindered this movement included the initial lack of validated software, the (realistic) fear of complexity, realistic implications of using 'black box' technology and the perceived costs, either for off the shelf software or internal research and development costs. The lack of widespread acceptance of the need for change is serious and has potentially delayed research in this area. We discuss these barriers briefly and provide plausible solutions.

Validated software applying the drop model was reported in 2007 [15] but was not made available either commercially or as freeware. Perlin et al. has reported a validated and commercially available software, TrueAllele® [2], that has proceeded through the court process in the US and UK. More recently Haned and Gill [12] have developed open source software. These software clearly go a long way to solving the first barrier; the availability of software. In 2004 Buckleton et al. stated "Once reliable continuous methods become available the binary method will have to be viewed as "second best" and will become obsolete" [16]. This prediction has yet to be fully fulfilled and it is important for the forensic community to consider why this is so.

A fully continuous method employs standard statistical theory to utilise as much available quantitative information as possible within a DNA profile. As more analysis parameters are incorporated the model increases in complexity and becomes more computationally demanding. Without an understanding of the underlying mathematics, the risk is that these systems become 'black boxes' whose workings are not understood by or, if intellectual property is concerned, are not accessible to the user. Presentation of any statistical analysis in court becomes problematic. The solution is the development of comprehensible models within the reach of the average forensic scientist coupled with upskilling and support for those forensic scientists. The goal is that a forensic biologist may effectively represent the evidence in court.

In this paper we describe the development of a fully continuous method for the relatively quick interpretation of low level and mixed STR DNA profiles that utilises peak height information from the DNA profile. We investigate reproducibility and speed and compare the outputs with binary methods.

Using a continuous system to analyse DNA profiles has the additional advantage that the analysis results (by necessity) in a sampling from the posterior distribution of each parameter included in the model. This allows the scientist additional points of review, as they can scrutinise properties of the DNA profile against the results of the analysis.

We identify the requirements for a continuous model as availability, short run time, comprehensibility and rigour.

2. Method

2.1. Mathematics

We may consider the evidence of the crime stain G_C to consist of a vector of observed peak heights **O** made up of a number of individual observed peak heights O^l_{ar} for allele a at locus l for replicate r. Let there be R replicates and L loci.

To form the likelihood ratio we consider two hypotheses H_1 and H_2 chosen to align with those expected for the prosecution and the defence respectively. The person of interest (POI) is assumed to be present under H_1 but not under H_2. The assumption of the POI (or a number of POIs) donating under H_1 requires that all the sets defined under H_1 contain the genotype(s) of the POI. The sets defined under H_2 may or may not contain a genotype corresponding to the POI. Refer to Appendix B for further detail.

We seek the likelihood ratio (across all loci) $LR = \Pr(G_C|H_1)/\Pr(G_C|H_2)$. Let H_m specify the J sets of N genotypes $\{S_j : j = 1, ..., J\}$ then

$$Pr(G_C|H_m) = \sum_{j=1}^{J} Pr(G_C|S_j)Pr(S_j|H_m).$$

Introducing the genotype combinations S_j and noting that once the sets S_j under each hypothesis are specified the hypotheses themselves are not required:

$$LR_C = \frac{\sum_j Pr(G_c|S_j)Pr(S_j|H_1)}{\sum_{j'} Pr(G_c|S_{j'})Pr(S_{j'}|H_2)} \quad (1)$$

where LR_C stands for the continuous LR and we make the point that the number of sets under H_1 may be very different from those under H_2 by using different subscripts. It is helpful to write $Pr(G_c|S_j) = w_j$ which may be usefully thought of as a goodness of fit of the data (G_C) to the genotype set S_j. Peak information from multiple replicates may be utilised in the generation of weightings for a genotype set.

$$LR_C = \frac{\sum_j w_j Pr(S_j|H_1)}{\sum_{j'} w_{j'} Pr(S_{j'}|H_2)} \quad (2)$$

More typically in most forensic calculations $S_{1,...,J'}$, specified by H_2, will contain the $S_{1,...,J}$ sets specified by H_1 and may contain additional genotype sets. As the weight is independent of the hypothesis given the genotype set, Eq. (2) may be evaluated as long as we have a complete set of weights for all required genotype sets, S_j and $S_{j'}$.

A method in common usage, the binary model, assigns the terms w_j the value 0 or 1 depending on whether the crime profile is deemed impossible or possible if it originated from the genotypes specified. As such it makes partial use of the information present in the peak height information.

Since the probabilities $\Pr(S_j|H_m)$, which represent genotype frequencies or probabilities, have been discussed extensively and there are accepted methods it would be sufficient and beneficial to solve for w_j and $w_{j'}$. We assert that this is beneficial because these fits of the data to the genotype set may be visualised by experienced forensic biologists and benchmarked against their judgement. This, we suggest, is a key to comprehensibility, a view we believe we share with Perlin et al. [2].

We introduce parameters to describe the true template level. Experience and empirical studies suggest that the height of peaks from a single contributor are approximately constant across the profile but generally have a downtrend with increasing molecular weight. Given this general downtrend individual loci may still be above or below the trend. In addition, the slope of the downtrend trend may vary from one contributor to another. The product from the amplification of an allele is dominated by correct copies at the allelic position and backstutter at one repeat shorter than the allele. There are a number of other more minor products ignored in this treatment. We term the sum of the allelic and backstutter product as total allelic product. We require a term for the true but unknown template level available at a locus for amplification. This is a function of the input DNA and any degradation or inhibition

effects. Since template is described by weight, usually in picograms, we coin the term mass to subsume the concepts of template, degradation, inhibition and any other effect that determines the expected total allelic product at a locus.

Hence, the mass of an allele at a locus is modelled as a function of various parameters which we collectively term the mass parameters. These are:

1. A constant t_n, for each of the n contributors that may usefully be thought of as template amount.
2. A constant d_n, which models the decay with respect to molecular weight (m) in template for each of the contributors to genotype set S_j. This may usefully be thought of as a measure of degradation.
3. A locus offset at each locus, A^l to allow for the observed amplification levels of each locus. We set $A^l = 1.00$ for one locus selected to be that locus with the largest total fluorescence. Note that individual A^l values may exceed 1.00 if that locus is above the general degradation trend.
4. A replicate multiplier R_r. This effectively scales all peaks up or down between replicates. We set $R_1 = 1.00$.

The total allelic product, T^l_{anr}, for an allele, a, at locus l, from contributor n in replicate r is modelled as:

$$T^l_{anr} = A^l R_r t_n e^{d_n \times m^l_a} X^l_{an} \tag{3}$$

where m^l_a is the molecular weight of allele a at locus l and X^l_{an} is the count of allele a at locus l in contributor n. The terms t_n and d_n collectively model an exponential decay of amplifiable template. All degradation values are constrained to be negative. $X^l_{an} = 1$ for a heterozygote with a and $X^l_{an} = 2$ for a homozygote a. The total allelic product from an allele is split between backstutter and allelic peak. We model stutter ratio π^l_a for allele a at locus l using a simplified model which assumes stutter height to be linearly proportional to allele height. The height of the allelic and stutter peaks formed from allele a are therefore:

$$E^l_{anr} = \frac{T^l_{anr}}{1 + \pi^l_a}$$

$$E^l_{(a-1)nr} = \frac{\pi^l_a T^l_{anr}}{1 + \pi^l_a}$$

where $a - 1$ signifies the stutter product.

The contributions from each contributor allele and stutter are then summed to produce a vector of expected peak heights, **E**, given a set of mass parameters. The variance for the logarithm of a combined allelic and stutter peak is assumed to follow the same form as an allelic peak of the same height as the sum.

We write the mass variables, $\{d_n : n = 1, \ldots, N\}$, $\{t_n : n = 1, \ldots, N\}$ as **D** and **T** respectively, $\left\{A^l : l = 1, \ldots, L\right\}$ as **A** and $\{R_n : n = 1, \ldots, R\}$ as **R**. The variables **D**, **A**, **R** and **T** are written collectively as **M**. Note that A^l for one locus and R_r for one replicate are fixed to 1.

We seek the integral across these nuisance parameters: $w_j = \int_M Pr(O^l_{11} \ldots O^l_{AR} | S_j, M) Pr(M) dM$. We assume w_j as the product across loci of the weights at a locus $w_j = \prod_l w^l_j$ and the weight at a locus as the probability of the peak heights across all replicates given the mass: $w^l_j = \int_M Pr(O^l_{11} \ldots O^l_{AR} | S_j, M) dM$. We constrain d_n to be negative but otherwise use uninformative priors, $\Pr(t_n)$ and $\Pr(d_n)$ for each of the N contributors, and the $R - 1$ priors, $\Pr(R_r)$. The probability of each of the $L - 1$ locus specific amplification efficiency (A^l) parameters for each of the $L - 1$ loci is modelled as $N(\mu_A, \sigma_A)$ where μ_A is the simple arithmetic average of the A^l values and σ_A is a preset hypervariable. This allows a limited freedom to the **A** variables but penalises any single value that departs significantly from the average. We use an uninformative prior on μ_A. Since we set one A^l term at 1.00 and there is a penalty for deviation from the average these collectively work to keep the A^l values close to each other and to 1. This treatment models the mass at each locus as generally flat or downwards with respect to molecular weight and correlated across loci. Each locus is permitted some limited freedom to deviate from the trend.

We have attempted to keep the number of free variables describing mass low. In total we have $2N + R$ free variables describing DNA mass. For example, for one replicate of a two person mixture this is 5. We suggest that, if we model mass correctly then the difference between observed and expected T^l_{anr} (total allelic product) for the alleles contributing to a profile are uncorrelated. We further suggest that this is diagnostic, if the difference between observed and expected T^l_{anr} values are found to be uncorrelated then we have modelled mass correctly.

Assumption 1. Peak heights are assumed to be conditionally independent given S_j and M:

$$Pr(O^l_{11} \ldots O^l_{AR} | S_j, M) = \prod_a \prod_r Pr(O^l_{ar} | S_j, M)$$

Our model is therefore $w_j = \int_M \prod_l \prod_a \prod_r Pr(O^l_{ar} | S_j, M) Pr(M) dM$ (4)

Assumption 1 is unlikely to be true, most obviously a larger than expected stutter peak is expected to be associated with a smaller than expected allelic peak. However data analysis suggests that the correlation is small or non-existent [17].

At steady state a Markov Chain using the Metropolis–Hastings algorithm is expected to spend time on a genotype set proportional to the probability of the profile given that genotype set. We obtain the weights w^l_j of each genotype set S^l_j, at each locus l as the fraction of MCMC iterations involving S^l_j.

In taking the weight across all loci as the product of the weights at each locus we have made a fairly standard assumption of independence of weightings between genotype sets of different loci, which is unlikely to be exactly true, and we investigate in Section 3. This decision was made because the weights at a locus can be directly visualised, but multi-locus weights quickly become incomprehensible to an analyst.

Assumption 2. The weight across a profile is the product of the weights at each locus

$$w_j = \prod_l w^l_j$$

which gives:

$$w_j = \prod_l \int_M \prod_a \prod_r Pr(O^l_{ar} | S_j, M) Pr(M) dM \tag{5}$$

as an approximation for Eq. (4). Because of the way we run the MCMC chain the variation in **M** is small and therefore the error in the substitution above is expected to be small. We discuss the performance of this, and other, approximations in Section 3.

For each trial of the current state and a proposed state, one or other of these states is accepted. This is often termed an iteration using the analogy of a Markov Chain as a random walk. However the actual action may be to move or stay still. We monitor how extensively the space has been searched and how efficient mixing has been and terminate the chain after a predetermined number of

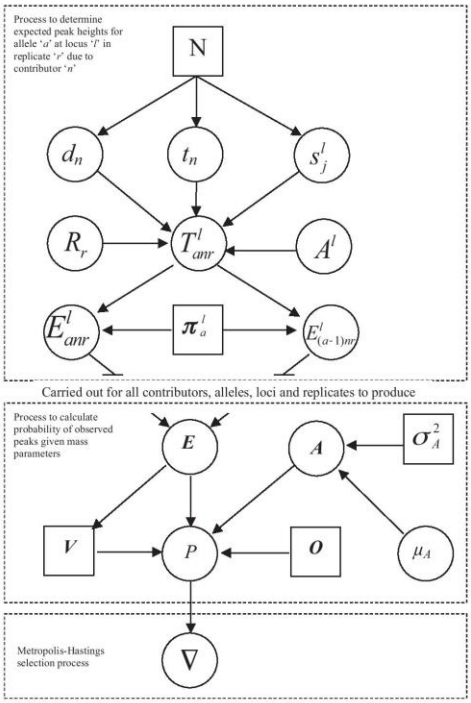


Fig. 1. A network diagram showing the functioning of the continuous model described here. Squares represent fixed quantities that are not optimised by the MCMC. The dashed box surrounds separate processes.

moves to new states, an approximate measure of how extensively the space has been explored.

2.2. Practical implementation of the mathematics

The applied model is represented as a network diagram in Fig. 1. The number of contributors N is assigned by the analyst after an assessment is made of the DNA profile. There is no mathematical need to constrain the number of contributors, N, but we add a practical constraint that $N = 1–4$.

The probability of obtaining the observed profile(s) given the mass parameters is calculated by considering the ratio of observed (O^l_{ar}) and expected peak heights following a model suggested by empirical data [17]. If a peak has been observed (i.e. $O^l_{ar} \geq Z$, where Z is the analytical threshold) we model $\log(O^l_{ar}/E^l_{ar})$ as normally distributed with mean 0 and variance c^2/E^l_{ar}:

$$Pr(O^l_{ar}|E^l_{ar}) = Pr\left[\log\left(\frac{O^l_{ar}}{E^l_{ar}}\right)\right] \quad \text{where} \quad \log\left(\frac{O^l_{ar}}{E^l_{ar}}\right) \sim N\left(0, \frac{c^2}{E^l_{ar}}\right) \qquad (6)$$

When the expected peak is not observed, we calculate $\int_{\lambda=0}^{Z} Pr(\lambda) Pr[\log(\lambda/E^l_{ar})]d\lambda$ where $\log(\lambda/E^l_{ar})$ is modelled as normally distributed with mean 0 and variance c^2/E^l_{ar}:

$$Pr(O^l_{ar} < Z|E^l_{ar}) = \int_{\lambda=0}^{Z} Pr(\lambda) Pr\left[\log\left(\frac{\lambda}{E^l_{ar}}\right)\right] d\lambda \qquad (7)$$

$$\text{where} \quad \log\left(\frac{\lambda}{E^l_{ar}}\right) \sim N\left(0, \frac{c^2}{E^l_{ar}}\right)$$

The variance is modelled as inversely proportional to the expected peak height. This model is informed from empirical studies separate from this paper [17,18]. We typically train the two external parameters on validation data usually a set of ten single source genotypes at ten dilutions down to levels were dropout occurs. The software is in casework use in Western Australia, South Australia, New South Wales, Queensland, Victoria and New Zealand (using Applied Biosystems' Profiler Plus®, Identifiler™, MiniFiler™, SGMPlus™ (at 34 cycles) multiplexes and Promega's PowerPlex® 21 multiplex) and under trial in the remaining Australasian jurisdictions (using Promega's PowerPlex® 21 multiplex and Applied Biosystems' 3130 and 3500 capillary electrophoresis instruments). For labs using the same multiplex and the 3130 instrument differences in the parameters were small. This suggests that minor changes in protocol may not require full revalidation. Equation (7) models the probability of dropout. Since we allow, but do not require, a threshold it is necessary to model the probability of a peak anywhere below the threshold. Dropout probabilities are applied to each expected, but unobserved, peak that is specified by the 'catch all' allele 'Q'. In doing so we make the approximation that multiple Q alleles are not the same. We depart from the logistical regression approach given by Tvedebrink et al. [19,20] but obtain a consistency with our other modelling in this paper.

The profile is pre-processed into a vector of peaks, compiled from peaks observed in evidence profile replicates. Peak data are edited in two ways.

First, any peak below an optional analytical threshold is deemed unobserved. This step is mathematically unnecessary and potentially counterproductive but comports with user requirements. Artefacts such as pull-up and peaks in forward stutter positions, but not in backstutter positions, are removed manually prior to interpretation. The removal of artefacts in allelic positions, such as potential forward stutter, leads to a weakness in our approach. At this stage this is unavoidable until we develop a probabilistic model for forward stutter.

The 'catch all' allele, Q, (or multiple Q alleles if required) is also included in this vector and represents all alleles not present in the observed EPG(s). This facilitates treatment of dropout for each locus. An exhaustive set of genotype combinations corresponding to the N contributors specified prior to processing is compiled using the vector of alleles.

If a contributor is assumed to be of a known genotype, say the complainant, under all H_m hypotheses the genotype of this contributor is fixed in the genotype sets. For some proposed genotype sets drop-in may be necessary to explain the observed EPG(s). At this stage we have a flexible but arbitrary model for the probability of the height of a drop in peak. Drop-in events are assigned probabilities based on the formula $\alpha e^{-\beta O}$ where α and β are constants. If $\beta = 0$ then the probability of a drop-in event (regardless of height, O) is a constant α. Otherwise the values can be set to $\alpha = \beta$ to assign probabilities based on observed peak height.

The complete list of genotype sets quickly increases in length with number of contributors and profile complexity. For a locus in which 'a' alleles have been detected there are $(a+1)^{2N}$ genotype sets initially considered. This is 'a' alleles plus a wildcard for dropout that can occupy each of the two allelic positions of each contributor. This is usefully reduced to a set of reasonable genotype vectors using three heuristics:

1) A pre-defined amount of allowable drop-in per locus, and
2) A pre-defined maximum stutter ratio, and
3) Removal of duplicated profiles.

After this screening process a list of J reasonable genotype sets is obtained.

We then implement component wise MCMC using the Metropolis–Hastings algorithm.

A vector of values for **M** is chosen from the range of possible values which is updated during the chain. The range of each element in **T** (the template of each contributor) is initially constrained by 0 and the amount of DNA required to explain the highest observed peak. **A** is initially constrained to be between 0.95 and 1.00 and **R** is similarly constrained to a window of width 0.05.

Since many of our priors are currently uninformative and only relative probabilities are required for the Metropolis–Hastings algorithm we may drop them from the calculation.

$w_j = \prod_l \prod_a \prod_r Pr(O^l_{ar}|S^l_j, M)Pr(A)$ is calculated for iteration 1, hereafter called Pr(1), by comparing expected peak heights to observed peak heights in each of the R replicates, where S^l_j is the genotype set j that is currently chosen for locus l.

With each MCMC step one side of the range of the elements within **S** and **T** are reduced by a predefined percentage of the current range, down to a minimum window, again set by a predefined value. The range for elements in **A** starts as 0.95–1.00, with each A^l chosen at random from within this range. The locus with the greatest observed total peak height (sum of peak heights for all observed peaks) is anchored to $A^l = 1.00$ and the window for all other loci (maintaining the 0.05 window) are moved down or up if MCMC accepts contain locus specific amplification efficiency (A^l) values at the extreme 0.01 of the range. The starting mid-point of the range of **R** is determined by the total peak height observed over a profile divided by the total peak height of the first replicate. This will set $R_1 = 1$ and for the entirety of the MCMC is fixed as 1. Again the R_r window for other replicates slides in the direction of a value in the extreme 0.01 of the range with each MCMC step.

For each new proposed state a locus l is chosen at random and a genotype set for that locus S^l_j is chosen from the pre-processed list of reasonable genotypes. A vector of values for **M** is chosen from the reduced range. The genotype and parameter vectors in iteration y move if $\Pr(y) \geq \Pr(y-1)$ or if a randomly chosen value from $U[0,1] < Pr(y)/Pr(y-1)$.

As explained above, for each iteration of the MCMC chain, elements of **M**, and a genotype set (S_j) that differs at one locus l are chosen.

For the state of the MCMC after each step, a tally of the number of times S^l_j has featured is incremented. The tallies of the S^l_j are normalised at each locus for analyst comprehension (this step is mathematically unnecessary) and interpreted as the weights of that genotype set at that locus w^l_j.

An expected peak height E^l_{ar} is generated using the model based on that described in Section 2.1.

We currently implement a simplified stutter model that models stutter ratio as linear with respect to the allelic designation rather than longest uninterrupted sequence (*LUS*) [21,22]. This was chosen because:

1. There are some alleles where it seems likely that we currently have a wrong value for *LUS*, and
2. We have no *LUS* value for many rare alleles, and
3. We are uncertain whether currently available *LUS* data derived from largely African American and Caucasian samples translates simply to Maori, Polynesian and Australian Aboriginal samples because specific sequence data is not available.

Expected peak heights are assumed to be additive when there are multiple contributions to a peak, whether from multiple alleles or a combination of alleles and stutter.

Empirical data suggests that the variance in a stutter peak in a model based on *LUS* follows a different pattern to an allelic peak [17]. In general, stutter peaks show less relative variance than allelic peaks. Since we have not yet implemented a *LUS* based model it is inappropriate to apply this variance and a larger variance is required and applied. We currently implement the same variance for a stutter peak as for an allelic peak. This defers, temporarily, the problem of solving the correct variance for a combined peak.

It has proven necessary to apply a correction for overamplified profiles. At high O^l_{ar} the observed peak is smaller than predicted by Eq. (3). This is expected as the linear relationship between peak height and DNA template only exists over a certain range. The correction can be specified by the user in the form of a saturation cap. Above this cap peak heights are considered qualitatively. Where there is a large peak observed above the saturation cap, all expected peak heights above this level are given identical probabilities when compared to the observed height.

2.3. Calculation of an LR

2.3.1. Point estimate

A point estimate for the *LR* is assigned using equation 2 with weights w_j and $w_{j'}$ provided from Eq. (5). Note that if $S_j = S_{j'}$ then $w_j = w_{j'}$. $\Pr(S_j|H_1)$ and $\Pr(S_{j'}|H_2)$ are developed using the population genetic model described in Balding and Nichols [23] and assuming all reference profiles are involved in the conditioning. θ values are input by the user. Allele probabilities are assigned as $(x_i + 1/k)/(N_a + 1)$ where x_i is the number of observations of allele i in a database, N_a is the number of alleles in that database and k is the number of allele designations with non-zero observations in the database. This is consistent with our later offering for assessing sampling uncertainty.

2.3.2. Sampling variation

We offer the highest posterior density (HPD) method for accounting for sampling variation [24–26]. The HPD method (sometimes referred to as Bayesian credible intervals) uses a Dirichlet distribution to describe the frequencies of alleles in a database. The Dirichlet distribution is a multivariate probability density function which describes the probability of obtaining a set of allele frequencies given the allele count obtained from a database. The general form of a Dirichlet distribution is:

$$f(\alpha_1 \ldots \alpha_n) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} x_i^{\alpha_i - 1}$$

referred to as a Dirichlet $(\alpha_1 \ldots \alpha_n)$ distribution where α_i is the frequency of allele i, x_i is the count of allele i from the database and k is the number of alleles at the locus with non-zero count.

We utilise a $1/k$ prior distribution, i.e. a Dirichlet $(1/k, 1/k, \ldots, 1/k)$ distribution to produce a Dirichlet $(x_1 + 1/k, x_2 + 1/k, \ldots, x_k + 1/k)$ posterior density. To calculate an exact credible interval would require the derivation of a probability density function for each likelihood ratio formula based on the Dirichlet distributions. This is difficult even for simple equations and so a Monte Carlo simulation method can be used that generates sets of allele frequencies (or vectors of allele frequencies) from a Dirichlet probability density function and then use these allele frequencies to generate likelihood ratios.

A fast method to sample a random vector $f_1 \ldots f_k$ from the k-dimensional Dirichlet distribution with parameters $(\alpha_1 \ldots \alpha_k)$ is to draw k independent random samples $g_1 \ldots g_k$ from gamma distributions each with density $\text{Gamma}(\alpha_i, 1) = g_i^{\alpha_i - 1} e^{-g_i} / \Gamma(\alpha_i)$ and then set $f_i = g_i / \sum_{j=1}^{k} g_j$ (so that the sum of all f_i frequencies is 1).

Gamma number generation was carried out using the GS and GD algorithms of Ahrens and Dieter [27,28]. We recognise credible

arguments that assessing uncertainty from sampling is unhelpful. Of the many valid arguments we have heard the one of greatest impact on us is that it may give the false impression that all uncertainty has been assessed.

We have implemented these methodologies in Java and have created software to examine the continuous approach to DNA profile interpretation. We suggest that the posterior distribution may be used to inform an assignment of the *LR*. If, say, the lower bound, is taken then if should be borne in mind that a statement that the *LR* is greater than this bound 0.99 of the time is only true if the only source of uncertainty is the sampling uncertainty. It may be better to accept that there are a great many sources of uncertainty and to assign the *LR* at this (or some other bound) as a subjective judgement of a fair and reasonable approach.

3. Tests of assumption 2 and non-thinning

The component wise MCMC chain varies one locus of the genotype at a time. This increases the MCMC mixing after burn-in by avoiding examining very unlikely genotypes too often. In addition we have assumed $w_j = \prod_l w_j^l$ (assumption 2). We have tested this assumption and the component wise approach on three mixtures described as a mixture of two contributors both of good template (strong/strong), one good and one low template (strong/weak) and both low template (weak/weak). We trialled the proposed approximation termed unthinned product, against our gold standard attempt at the optimal solution. The gold standard was set as the thinned result of the full multilocus acceptances for a considerably extended MCMC chain thinned to record the accepted state after every 10th step. We plot $\log w_j$ for a random sample of 200 of the accepted genotypes. In Fig. 2 below we give the results.

The minimum possible value for the gold standard log weighting is −5.3 and occurs when there is one acceptance for this multilocus genotype. Since there is only one acceptance this value is poorly determined. It is certainly possible that the unthinned product is actually a better estimate of these low weightings. At the high weighting end, the top right of each graphic, the correlation between gold standard and unthinned product is good. The correlation in weightings produced by Eq. (4) (gold standard) and Eq. (5) (product of locus weightings) provides an empirical justification for making assumption 2.

4. Reproducibility

An issue with using a stochastic system like MCMC is that the results of no two analyses will be completely the same. This is an issue that is relatively new to forensic science although this is a feature of both TrueAllele and LikeLTD [29]. Up until this point forensic science has always had the luxury of, at least theoretically, completely reproducible results. This variation has troubled users.

Increasing the number of MCMC steps ameliorates but does not remove the variation. There is, however, an associated runtime cost. Hence a trade-off between reproducibility and runtime must be struck.

In the *LR* the numerator is the weighted sum of the probability of the data given fewer genotype sets than the denominator. In many cases the numerator may have only one term. Since the denominator is the weighted sum across the probability of many genotype sets it has a stability to variation in the weights. However the numerator is more sensitive and this effect is at its greatest when the weight for the numerator genotype set(s) is low.

To demonstrate reproducibility, a two person mixture was analysed 10 times under our standard running conditions. Average runtime for each of these analyses was 25 min on a standard desktop computer, however improvements since the time this experiment was run has seen the runtimes for the same problem drop to 2 min.

Contributor 1 produced an average *LR* of 4.27×10^{19} with a standard deviation of 2.7% of the mean. Contributor 2 produced an average *LR* of 4.6×10^{19} with a standard deviation of 2.9% of the mean. Likelihood ratios obtained from separate runs were highly reproducible. It is expected that the most variation in results would be seen when the likelihood ratio at one or more loci strongly favoured exclusion, due to poor fit of the observed data to the proposed contributors.

5. Validation experiments

Although the method described here was necessarily programmed as software we emphasise validation of the method as opposed to validation of the software. Indeed the software has been tested against by hand calculations but of more general interest would be the validity of the approach.

The method produces an *LR*. No true *LR* is available and many would argue no such thing exists. It is therefore not possible to examine the results against some true answer. The only practical tests that can be done are:

1. Examination of interpretations of mixtures of known contributors (ground truth) and
2. Comparisons against other methods and/or human judgement.

Neither of these is trivial. Ground truth comparisons should produce a large *LR* when a true contributor is tested and a low one when a false one is tested. However it is also reasonable that occasionally a known contributor could give a low *LR* especially when the profile is low template. Equally occasionally a false

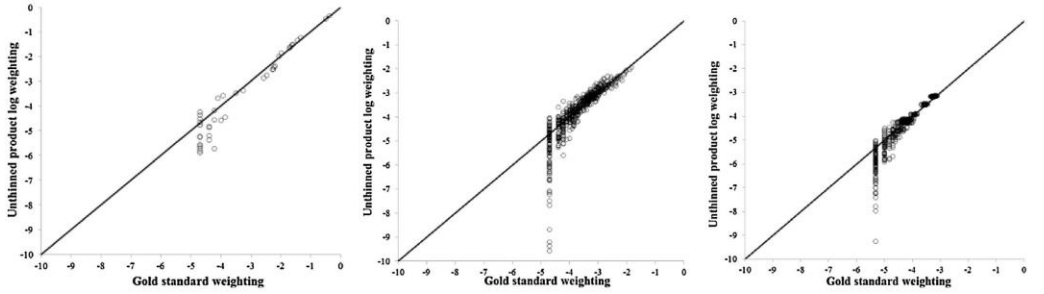


Fig. 2. A plot of the log weightings for the gold standard and the unthinned product for the strong/strong, strong/weak and weak/weak mixtures. The diagonal line represents the points at which the log weighting and gold standard weighting are equal.

contributor may give a high *LR*. This is termed an adventitious match. Such tests give little guidance as to whether the *LR* is too large or not large enough. Early versions of the model were beta tested by a number of users often using a set of two and three person mixtures where the ground truth was known. In the vast majority of trials undertaken a subjectively judged fair and reasonable *LR* resulted. All counter examples were examined and corrective action taken if necessary.

In Appendix C we give the results of 127 trials where H_1 is true. The expected result is a large *LR* and this occurred in all cases.

Comparisons were also undertaken against the binary model of Kelly et al. [30]. The method of Kelly et al. [30] is known to be wasteful of information hence it is not obvious what the expected result is. In all cases inspection of the results suggested that the *LR* assigned by the method was fair and reasonable (see Appendix C).

Comparisons against human judgement were undertaken as part of a formal trial (see Appendix C). All cases of counterintuitive results have been examined and corrective action taken if needed.

We hope that the method, to some extent at least, puts out diagnostics that allow an operator to "validate" the specific result. The weights for each genotype set are an intuitive thing that experienced operators can check. We would suggest that any strongly counterintuitive result warrants investigation and would strongly counsel against reporting results from any software without some form of human check.

6. Limitations of the system

We acknowledge some limitations of the proposed method. In broad terms the method described here is standard mathematics coupled with a model for peak height and the variability in peak height. The limitations arise in some cases from conscious choice and in others from the current state of model development.

The most obvious limitation we are aware occurs when a large artefact is allowed through the manual EPG review process. Peaks present in the EPG must be accounted for in the model by some biological phenomenon. An artefact peak on the EPG will be considered as an allele, or a drop-in, both of which can substantially shift the distribution of probability across genotype sets.

Currently the user must specify the number of contributors to a mixture prior to analysis and this number must be the same in both the numerator and denominator. For the future, we can envisage a system where the number of contributors is treated as an unknown nuisance parameter and integrated out and we intend to experiment with this.

We have implemented a stutter model which we know to be suboptimal as it uses allele designation rather that longest uninterrupted sequence of repeats. In extensive trials there appears to be little consequence of this however we do intend to update the model in future revisions.

We use a catch-all allele designation '*Q*' for all alleles that have not been observed in G_C replicates. We consider separate *Q*'s for each dropped allele in the genotype set, each with their own dropout probability and E^l_{an}. The use of the *Q* designation shortens the runtime as it reduces the number of alleles under consideration by the MCMC, however represents a small loss of accuracy. All alleles being considered under a *Q* designation will be given an average allele molecular weight for that locus and Q alleles that overlap are not summed. The choice to do this was based on user requirements.

7. Conclusion

Our ability to generate DNA profile has far outreached our ability to interpret them. Binary methods of DNA interpretations, which rely on a rule and threshold based system have been pushed to their limits, and it seems that the next step for DNA profile interpretation is to move to continuous systems. Continuous systems can utilise all the currently observed information within a DNA profile.

The superiority of continuous systems has been acknowledged for over a decade but implementation has lagged. We have speculated on the causes and highlight the lack of validated software, cost, runtime but perhaps most significantly a concern about black boxes and the inability of analysts to effectively represent the evidence in court. The acknowledged success of the TrueAllele® software in casework and court has begun to erode these barriers but uptake would still represent a minority of casework.

We report here a synergy of mathematics and biological modelling specifically designed to ameliorate the black box aspect of continuous models. Teaching of MCMC methods and the use of the software to forensic biologists has been successful with participants having positive reactions. Run times and reproducibility are acceptable.

It is difficult to test such systems for the accuracy of the *LR* produced. The correct answer is unknown and in many cases unknowable. We resist the temptation to state that big numbers are good.

We have judged the results against the known input contributors. We score a result as good if it produced a large *LR* for a correct hypothesis and a low *LR* for a false hypothesis. Further each genotype combination may be examined with regard to the weighting assigned. Combinations with high weight subjectively provide the best explanation of the data with impressive regularity as seen in trial comparing results of human assessments with those of the continuous model outlined in this paper. Against this criterion, which we acknowledge as subjective, results are outstanding.

Acknowledgements

We would like to thank James Curran (University of Auckland), Steven Myers (California Department of Justice), Michael Coble (NIST), Peter Green (University of Bristol) and Todd Bille and Steven Weitz (Bureau of Alcohol, Tobacco, Firearms and Explosives) for their helpful discussions and contributions during the development of this work. We would also like to the SallyAnn Harbison for her helpful comments that improved this paper. The authors gratefully acknowledge the source of some of the data that lies behind Appendix C. These data were provided by the state and territory government forensic biology laboratories of Australia. Finally we would like to thank two anonymous reviewers, whose comments improved this work. This work was supported in part by grant 2011-DN-BX-K541 from the US National Institute of Justice. Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the U.S. Department of Justice.

Appendix A. A summary of nomenclature used within this paper

- a the allele
- $a-1$ signifies the stutter product for allele a
- **A** the mass variable for locus amplification efficiency, $\left\{A^l : l = 1, ..., L\right\}$ – locus offset at locus l.
- c a constant in modelling the variance in peak height
- **D** the mass variable for degradation, $\{d_n : n = 1, ..., N\}$ – degradation in template vs. molecular weight for contributor n
- **E** the vector of expected peak heights
- $E^l_{anr} = T^l_{anr}/(1 + \pi^l_a)$ the contribution of contributor n to the expected height of the allelic peaks at locus l formed from allele a in replicate r

- $E^l_{(a-1)nr} = \pi^l_a(T^t_{anr})/(1+\pi^l_a)$ the contribution of contributor n to expected height of the stutter peaks at locus l formed from allele a where $a-1$ signifies the stutter product in replicate r
- G_C the evidence of the crime stain across all R replicates
- H_m hypotheses, H_1 and H_2 hypotheses chosen to align with the prosecution and the defence, respectively
- J the number of contributors with j representing a specific contributor
- L the number of loci with l representing a specific locus
- LR_C the continuous LR
- LR_B the binary LR
- LUS the longest uninterrupted sequence within an allele
- **M** is the mass variables **D**, **A**, **R** and **T** collectively
- m^l_a is the molecular weight of allele a at locus l
- N number of contributors with n representing a specific contributor
- **O** the vector of observed peak heights
- O^l_{ar} the observed peak height for allele a at locus l for replicate r
- P the probability of observed data given mass parameters
- Q a catch-all allele to cover all possibilities outside a specified set
- R number of replicates with r representing a specific replicate
- **R** the mass variable for replicate amplification, $\{R_r : r = 1, ..., R\}$ is a multiplier applied to replicate r. We constrain $R_1 = 1$
- S_j the N contributor genotype set j
- S_k the genotypes of known contributors under both H_1 and H_2
- S_p genotype of the person of interest (a known contributor under H_1 only)
- S_u the genotypes of unknown individuals in S_j required under H_m
- **T** the mass variable for template DNA, $\{t_n : n = 1, ..., N\}$ – template DNA for contributor n
- $T^t_{anr} = A^l R_r t_n e^{d_n \times m^l_a} X^l_{an}$
- X^l_{an} is the count of allele a at locus l in contributor n. $X^l_{an} = 1$ for a heterozygote with a and $X^l_{an} = 2$ for a homozygote a
- π^l_a stutter ratio for allele a at locus l
- **V** the variance variables c and σ^2_A
- **V** the vector of all variance values V_{ar} given the expected peak heights, **E**, and the variance constant, c
- V_{ar} Variance of peak a in replicate r
- w_j the weight for genotype set j, $w_j = Pr(G_c|S_j)$
- w^l_j the weight for genotype set j at locus l
- Z analytical threshold below which peaks are deemed unobserved
- σ^2_A hyper-parameter for variance of **A**
- μ_A hyper-parameter for mean of **A**
- $\bigtriangledown$ Metropolis–Hastings acceptance/rejection algorithm for mass parameters

Appendix B. Derivation of *LR* and use of MCMC to determine weightings

The LR seeks to calculate:

$$LR = \frac{Pr(E|H_1)}{Pr(E|H_2)}$$

H_1 and H_2 specify sets of genotypes for consideration. Including genotype sets in the *LR* by the law of 'extending the conservation' gives:

$$LR = \frac{\sum_j Pr(E|S_j)Pr(S_j|H_1)}{\sum_j Pr(E|S_j)Pr(S_j|H_2)}$$

Note:

- The $Pr(E|S_j)$ does not depend on the hypothesis and so the hypotheses can be removed from the conditioned terms. This is because the hypothesis denotes genotypes sets, so both are not required.
- We have not discriminated between genotype sets in the numerator and denominator. This is because the terms $(E|S_j)$ are independent of hypothesis and it is terms of the form $Pr(S_j|H_m)$ that are conditional on the hypotheses.
- This can be shown by transforming S_j to S_u (for unknown genotype set u) and conditioning on S_k (the genotypes of known contributors under H_m) where $S_k \cap S_u = \varnothing$ and $S_k \cup S_u = S_j$.

$$LR = \frac{\sum_u Pr(E|S_j)Pr(S_u|S_{k^*}, H_1)Pr(S_{k^*}|H_1)}{\sum_u Pr(E|S_j)Pr(S_u|S_k, H_2)Pr(S_k|H_2)}$$

Note that $S_{k^*} \neq S_k$ as the known contributors under H_1 may differ from those known under H_2.

We can split S_{k^*} further into S_p (the known contributors under H_1, but not under H_2) and S_k (the known contributors that are known under both H_1 and H_2):

$$LR = \frac{\sum_u Pr(E|S_j)Pr(S_u|S_k, S_p, H_1)Pr(S_k|S_p, H_1)Pr(S_p|H_1)}{\sum_u Pr(E|S_j)Pr(S_u|S_k, H_2)Pr(S_k|H_2)}$$

Now removing independent terms:

$$LR = \frac{\sum_u Pr(E|S_j)Pr(S_u|S_k, S_p, H_1)Pr(S_k)Pr(S_p|H_1)}{\sum_u Pr(E|S_j)Pr(S_u|S_k, H_2)Pr(S_k)}$$

And cancelling $Pr(S_k)$ terms:

$$LR = \frac{\sum_u Pr(E|S_j)Pr(S_u|S_k, S_p, H_1)Pr(S_p|H_1)}{\sum_u Pr(E|S_j)Pr(S_u|S_k, H_2)}$$

Finally if the genotype of the person of interest is a known under H_1 then $Pr(S_p|H_1) = 1$ giving:

$$LR = \frac{\sum_u Pr(E|S_j)Pr(S_u|S_k, S_p, H_1)}{\sum_u Pr(E|S_j)Pr(S_u|S_k, H_2)}$$

Note that when there are no known contributors under both H_1 and H_2 then S_k can be an empty set $S_k = \varnothing$ and under this circumstance $S_j = S_u$. Similarly if the genotypes in S_p describe genotypes for all contributors then $S_u = \varnothing$ meaning that $Pr(S_u|S_k, S_p, H_1) = Pr(\varnothing\,|S_k, S_p, H_1) = 1$ and $S_j = S_p$.

To this equation we can consider a number of parameters that are used to describe a DNA profile, we have termed these mass parameters (**M**) and includes a template DNA amount for each contributor, a degradation level for each contributor, a locus amplification efficiency for each locus and a replicate amplification efficiency for each replicate. Including these nuisance parameters in the *LR* gives:

$$LR = \frac{\sum_u Pr(E|S_j, \mathbf{M})Pr(S_u|S_k, S_p, H_1)Pr(\mathbf{M})}{\sum_u Pr(E|S_j, \mathbf{M})Pr(S_u|S_k, H_2)Pr(\mathbf{M})} \quad (8)$$

This gives the exact form of the *LR* if we knew what all the mass parameter point values were. We do not know these values but can integrate across all possible values to give:

$$LR = \frac{\int_M \sum_u Pr(E|S_j, \mathbf{M})Pr(S_u|S_k, S_p, H_1)Pr(\mathbf{M})d\mathbf{M}}{\int_M \sum_u Pr(E|S_j, \mathbf{M})Pr(S_u|S_k, H_2)Pr(\mathbf{M})d\mathbf{M}}$$

which is the form of the *LR* given in the body of the paper in Eq. (2). This integral (which, if **M** was expanded out into individual parameters, is actually a multidimensional integral) is a problem that can be assessed using MCMC.

From Eq. (8) a MCMC analysis could be run separately for numerator and denominator. Each would choose values for

M and the acceptance/rejection probability would be calculated by:

$$\sum_u Pr(E|S_j,\mathbf{M})Pr(S_u|S_k,S_p,H_1)Pr(\mathbf{M}) \quad \text{and}$$

$$\sum_u Pr(E|S_j,\mathbf{M})Pr(S_u|S_k,H_2)Pr(\mathbf{M})$$

for H_1 and H_2 chains respectively.

To do this would mean that for every comparison to a person of interest in a case, an additional MCMC analysis would be required, which has associated time and computer resource costs. There would also be a between chain comparison requirement that would necessitate the calculation of the marginal probabilities of each chain.

To overcome this, we use MCMC to assess a single probability that does not take the two competing hypotheses into account. We instead assess:

$$Pr(E|H)$$

where H in this instance is simply that the profile has originated from 'n' contributors of which 'k' have genotypes S_k if this is appropriate ($k < n$). We substitute H for $N = n$:

$$Pr(E|N=n,S_k)Pr(N=n)Pr(S_k)$$

Due to our assumption on the number of contributors $\Pr(N = n) = 1$. Also, since S_k is a set of assumed contributors, $\Pr(S_k) = 1$.

We introduce $\Im$ which includes **M** and the possible genotypes of contributors as nuisance parameters. For convenience we also drop the $N = n$ and S_k as the genotype sets included in $\Im$ are restricted to 'n' contributor scenarios where $S_k \subset S_j$.

$$Pr(E|\Im)Pr(\Im)$$

We also note that including genotype sets as a variable in the MCMC means that only one genotype set is the focus of the MCMC at any time and acceptance/rejection probabilities are based on the below formula (where the genotype set in current focus S_j has been split from $\Im$ purely to remind the reader of their presence):

$$Pr(E|S_j,\mathbf{M})Pr(S_j)Pr(\mathbf{M}) \quad \text{Noting that}: Pr(S_j)Pr(\mathbf{M}) = Pr(S_j,\mathbf{M})$$

Having a single genotype set as the focus of the MCMC and calculating the acceptance/rejection probability on a single genotype set, decreases the acceptance rate, however offsets this by being able to complete an iteration with much less calculation. In fact the time increase due to a lower acceptance rate is less than the speedup due to quicker calculation time as there can be multiple genotype sets that have similarly high posterior probabilities. The MCMC used to assess this probability takes samples from the posterior distributions of all parameters in the model:

$$Pr(\Im|E) = \frac{f(E|\Im)Pr(\Im)}{f(E)}$$

where $f(E)$ is the marginalisation constant. As this is constant for the fixed dimension problems being discussed it will cancel out on all terms within the *LR* calculation and the above formula can be displayed as the familiar relationship:

$$Pr(\Im|E) \propto f(E|\Im)Pr(\Im)$$

Again expanding $\Im$ and concentrating on genotype set gives the posterior probability:

$$Pr(\boldsymbol{S}_j|E,\mathbf{M}) \propto f(E|\boldsymbol{S}_j,\mathbf{M})Pr(\boldsymbol{S}_j)Pr(\mathbf{M})$$

The right hand side of the above equation (likelihood multiplied by prior) corresponds to the probability we wish to assess in Eq. (8), and so the posterior probabilities calculated by the MCMC can be used.

We make one further modification during the MCMC. We bias the genotype set prior by $[\Pr(S_j)]^{-1}$ i.e. the inverse of the genotype set prior probability. This has the effect that the posterior probability is biased by the same factor. This adjustment is allowed as the biased prior for each genotype set is constant and can be corrected for post MCMC. In effect we have taken samples from:

$$\frac{Pr(S_j|E,\mathbf{M})}{Pr(S_j)}$$

For each S_j within the model. The choice to do this has the advantage that allele frequencies, and hence a population, does not need to be specified in order for deconvolution to be carried out. The population, or indeed multiple populations, of interest can be chosen at a later time when an *LR* is required.

As before we can expand genotype set based on genotypes of known S_k and unknown S_u contributors. In this instance we do not have S_p as we are assessing equation *Z*, which has only one hypothesis. So for H_2 in Eq. (8) we can substitute:

$$Pr(E|S_j,\mathbf{M})Pr(S_u|S_k,H)Pr(\mathbf{M}) = Pr(S_j|E,\mathbf{M})$$

And for H_1 in Eq. (8):

$$Pr(E|S_j,\mathbf{M})Pr(S_u|S_k,S_p,H_1)Pr(\mathbf{M}) = \begin{cases} Pr(S_j|E,\mathbf{M}) & \text{when } S_p \in S_j \\ 0 & \text{otherwise} \end{cases}$$

Remembering that each S_j is a set of 'n' single person genotypes and S_p is a set of 'p' genotypes of the known contributors under H_1 ($p \leq n$) so that S_p is a subset of each S_j.

We correct for the bias introduced into the posterior during the MCMC by multiplying each of the j posterior elements by $\Pr(S_j)$ within the *LR*.

The posterior $\Pr(S_j|E, \mathbf{M})$ enumerated from the MCMC is a distribution. We use the mean of this distribution in the *LR*. The mean can be determined directly from the posterior distribution. Alternatively, because genotype sets can take only discrete, unordered values the posterior probability for genotype set j can be determine by residence time of genotype set j as the focus of the MCMC. In fact the residence time of S_j in the MCMC will be directly related to its probability as this is what the Metropolis–Hastings acceptance/rejection criteria are based on.

Appendix C. Validation experiments

Testing the model when H_1 is true

Experiment 1: 127 artificially constructed mixtures were created from two contributors of known genotype. The mixtures were amplified following the manufacturer's recommended protocols with Applied Biosystems' Identifiler™ multiplex (Life Technologies, Carlsbad, CA) on an Applied Biosystem's 9700 thermal cycler. Amplified DNA was separated on an Applied Biosystem's 3130 capillary electrophoresis instrument. Resulting DNA profiles were analysed using Applied Biosystem's GeneMapper ID v3.2. The target total template was varied from 100 to 500 pg and the ratio of contributors was varied from 1:1 to 5:1. An *LR* was calculated for the hypotheses:

H_1: the DNA came from P_1 and an unknown person.

H_2: the DNA came from two unknown people.

P_1 was taken as the minor contributor but note that the minor contributor may be half the DNA in a 1:1 mixture. In Fig. 3 we give the

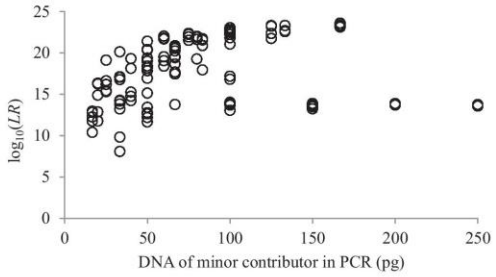


Fig. 3. A plot of *LR* vs template of the minor contributor for the ground truth H_1 true. The lowest template for the minor was 16.7 pg.

LR plotted against template of the minor contributor. *LR*s varied from 10^8 to the maximum possible 10^{23} and showed an increasing trend with increasing template. The highest values occurred for 2 or 3:1 mixtures at good template and the lowest values for mixtures of any ratio where the minor template was low. All the mixtures constructed to be 1:1 gave values in the range 10^{12}–10^{14} and represent the lower plateau of values in Fig. 3.

Fig. 3 demonstrates two phenomenons, which we would expect from a continuous system:

1) As the amount of template DNA decreases the *LR* decreases.
2) As the ability to resolve peaks as belonging to the major or minor contributor is lost the weightings for a number of genotype sets become equal. This 'spreading' of probability lowers the *LR* as probability is being shifted away from the genotype sets making up H_1.

It is worth noting that even at the lowest DNA amount of 16.7 pg, the continuous method as described in this paper produced a *LR* in the tens of billions.

Experiment 2: Artificial three- and two-person mixed DNA profiles with known contributors were amplified following the manufacturer's recommended protocols with Applied Biosystems' NGM SElect™ multiplex (Life technologies, Carlsbad, CA) on an Applied Biosystem's 9700 (silver block) thermal cycler. Amplified DNA was separated on an Applied Biosystem's 3130xl capillary electrophoresis instrument. Resulting DNA profiles were analysed using Applied Biosystem's GeneMapper ID v3.2.1.

The likelihood ratio was calculated for each contributor to each of the four two-person and six three-person mixed DNA profiles using both a binary method and the continuous method of interpretation discussed in this paper. The hypotheses considered are:

H_1: the DNA came from P_1 and unknown people up to the number of contributors.

H_2: the DNA came from all unknown people.

LR_B was calculated in MS EXCEL following the 'F model' as described in Kelly et al. [30].

Table 1 shows the likelihood ratio produced from the continuous method as described above and implemented through Java software, and a binary method for the same set of mixtures of known source calculated in MS EXCEL following the 'F model' described in Kelly et al. [30]. Profiles were of reasonable quality to allow assessment by the binary method. Table 1 shows the information gain by using a continuous system. In the two person scenarios where the individual profiles can be well resolved the information obtained from the two methods are similar. For three person mixtures, the results of the two systems diverged. LR_B was markedly lower or unable to be determined for three person mixtures, whereas LR_C continued to produce LRs consistently much greater than one.

The EPG for one of the three-person mixtures summarised in Table 1 (example six) in Fig. 4 below. For both the continuous and binary models the assumption of three contributors is made. This assumption is supported by inspection of Fig. 3.

Worked locus example

The workings for both *LR*s for the locus vWA are summarised in detail below for the hypotheses:

H_1: POI_2 and two unknown individuals.

H_2: three unknown individuals.

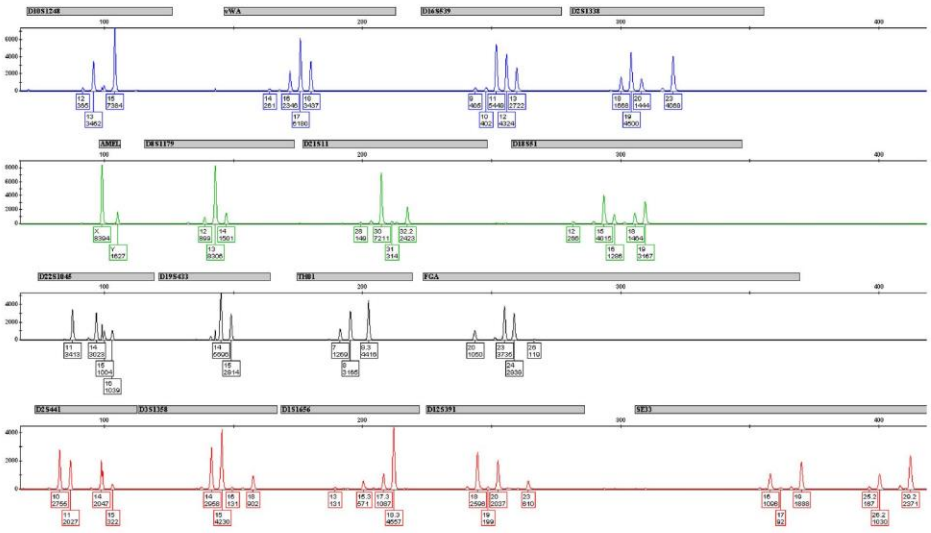


Fig. 4. EPG of three contributor mixed DNA profile.

Table 1
LR results of continuous versus binary method for assessing two and three person profiles, where NC stands for non-concordance (and therefore a statistic was not calculated).

Continuous		Binary	
Mixed DNA profiles from two contributors			
Person 1	Person 2	Person 1	Person 2
6.12E+15	3.16E+16	1.24E+16	3.81E+15
1.88E+19	6.37E+19	9.50E+17	1.93E+19
2.44E+19	4.87E+19	2.08E+18	9.61E+18
3.02E+17	1.25E+20	4.04E+14	1.25E+20

Continuous			Binary		
Mixed DNA profiles from three contributors					
Person 1	Person 2	Person 3	Person 1	Person 2	Person 3
2.72E+08	4.69E+11	3.79E+13	NC	226	1.39E+6
9.10E+08	4.17E+11	2.83E+13	NC	3	14,261
8.43E+19	4.73E+11	5.96E+18	57	<1	331
4.52E+12	5.73E+12	7.20E+14	<1	<1	846
9.62E+07	1.24E+20	1.07E+20	NC	65	317
7.68E+18	3.12E+19	1,254	23	356	NC

The reference DNA profiles of the three known contributors are:

Locus	POI_1	POI_2	POI_3
D10S1248	15,15	13,15	12,15
vWA	16,17	17,18	14,16
D16S539	13,13	11,12	9,10
D2S1338	18,20	19,23	20,23
D8S1179	13,14	13,13	12,13
D21S11	32.2,32.2	30,30	28,31
D18S51	16,18	15,19	12,15
D22S1045	15,16	11,14	11,16
D19S433	14,14	14,15	14,15
TH01	7,9.3	8,9.3	7,9.3
FGA	20,23	23,24	24,26
D2S441	11,11	10,14	14,15
D3S1358	15,18	14,15	15,16
D1S1656	15.3,17.3	18.3,18.3	13,16
D12S391	18,23	18,20	18,19
SE33	16,26.2	19,29.2	17,25.2

Allele frequencies were obtained from data analysis of an Australian Caucasian sub population ($N = 2082$, data not published) with the relevant allele frequencies for vWA below:

Allele	Allele frequency
14	0.1146
15	0.1071
16	0.2044
17	0.2726
18	0.2090

Continuous model of interpretation

After 50,000 MCMC accepts (10,000 of which were burn-in) the following weights (w_j) for each of the different genotype sets (S_u) using the models described above, were obtained for the one locus:

S_u	G_{C1}	G_{C2}	G_{C3}	w_j
S_1	16,18	17,17	14,14	0.00045
S_2	16,18	17,17	14,15	0.00017
S_3	16,18	17,17	14,16	0.00008
S_4	16,18	17,17	14,17	0.00002
S_5	16,18	17,17	14,18	0.00054
S_6	18,18	17,17	14,16	0.00005
S_7	16,16	17,18	14,14	0.00218
S_8	16,16	17,18	Q,14	0.00010
S_9	16,16	17,18	14,15	0.00207
S_{10}	16,16	17,18	14,16	0.00511
S_{11}	16,16	17,18	14,17	0.02030
S_{12}	16,16	17,18	14,18	0.00279
S_{13}	16,17	17,18	14,14	0.19300
S_{14}	16,17	17,18	Q,14	0.00368
S_{15}	16,17	17,18	14,15	0.15800
S_{16}	16,17	17,18	14,16	0.28700
S_{17}	16,17	17,18	14,17	0.21000
S_{18}	16,17	17,18	14,18	0.11400
S_{19}	17,17	17,18	14,16	0.00016

$Pr(E|H_1) = 3.87 \times 10^{-3}\ Pr(E|H_2) = 4.7 \times 10^{-4}\ LR_C = 8.24$

The mixture proportions determined by the software for the three individuals were 0.26, 0.70 and 0.04 for POI_1, POI_2, and POI_3, respectively.

The weightings list may be compared with human judgement as a diagnostic to check whether the software has operated reasonably. Note that set S_{16} has received the highest weighting and visual inspection of the EPG suggests that this is reasonable. It is also the set that corresponds to the three known contributors.

Binary model of interpretation

LR_B was calculated in MS EXCEL following the 'F model' as described in Kelly et al. [30]. The allelic vector for vWA was determined to be: 14,16,17,17,18,F.

$$\begin{aligned} Pr(E|H_1) &= 24Pr(14,16,17|14,16,16,17,17,18) \\ &= \frac{\begin{array}{c}24(\theta+(1-\theta)\times 0.1146)\times(2\theta+(1-\theta)\times 0.2044)\\ \times(2\theta+(1-\theta)\times 0.2726)\end{array}}{(1+5\theta)(1+6\theta)(1+7\theta)} \\ &= 0.1669 \end{aligned}$$

$$\begin{aligned} Pr(E|H_2) &= 360Pr(14,16,17,17,18|14,16,16,17,17,18)\times \\ &\quad 360(\theta+(1-\theta)\times 0.1146)\times(2\theta+(1-\theta)\times 0.2044) \\ &= \frac{\begin{array}{c}(2\theta+(1-\theta)\times 0.2726)\times(3\theta+(1-\theta)\times 0.2726)\\ \times(\theta+(1-\theta)\times 0.2090)\end{array}}{(1+5\theta)(1+6\theta)(1+7\theta)(1+8\theta)(1+9\theta)} \\ &= 0.1345 \end{aligned}$$

$LR_B = 1.24$

Summary of binary versus continuous profile interpretation

A summary of the calculated *LR* for all loci following the methods described above for the mixture in Fig. 4 is below:

Locus	LR_B	LR_C
D10S1248	0.97	4.69
vWA	1.24	8.21
D16S539	0.45	5.32
D2S1338	2.27	31.22
D8S1179	0.51	7.79
D21S11	0.94	9.98
D18S51	3.85	52.08
D22S1045	4.32	59.18
D19S433	0.92	7.17
TH01	0.97	13.31
FGA	1.39	21.14
D2S441	0.65	4.84
D3S1358	0.93	13.22
D1S1656	5.55	106.14
D12S391	1.42	21.34
SE33	6.23	69.53
Overall *LR*	356	3.12E+19

Experiment 3: Testing the model when the POI is not a donor

Of equal importance is the functioning of the model when the POI is not a donor. In the vast majority of scenarios where POI is not a donor of DNA to the sample, the *LR* is less than one. Fig. 5 shows the

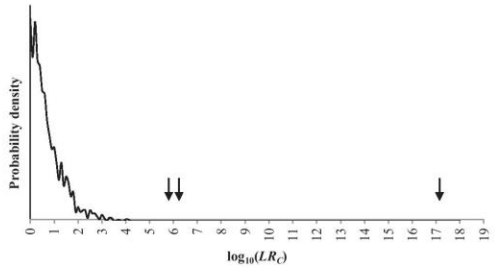


Fig. 5. Distribution of likelihood ratios (shown on a logarithmic scale) when considering known non-contributing individuals as contributors to a complex three person mixed DNA profile. Only values of $LR > 1$ are shown. Arrows show the LR for the three individuals, known to make up the mixture. The maximum value for the LR when comparing a known non-contributor was 12,896 ($\log_{10}(LR_C) = 4.1$).

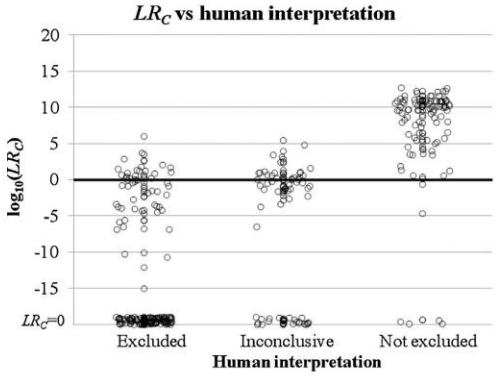


Fig. 6. Comparison of LRs produced using a continuous system to human interpretation. The line at 0 represents neutrality i.e. the probability of obtaining the profile is the same under hypotheses of exclusion or inclusion. Results above the zero line favour inclusion and results below the line favour exclusion. When $LR_C = 0$ (when G_P did not feature in the MCMC at any point) the result has been plotted at the bottom of the graph against the '$LR_C = 0$' y-axis label.

results of considering 57,612 individuals as potential contributors to a complex 3 person Identifiler™ profile. Of these 57,612 individuals, three were known contributors and 57,609 were known non-contributors.

Of the 57,612 individuals, only 1168 gave a LR in favour of H_1 and as Fig. 5 shows the majority of these were below a LR of 100. In contrast the known contributors gave LRs of greater than 700,000, and were clearly distinguishable from the non-contributors.

Trials against human interpretation

A continuous model for DNA interpretations should produce results that are intuitively correct to a trained scientist. We would therefore expect to see a relationship between LR_C and human interpretations.

To test this, previously reported casework Profiler Plus® profiles were reanalysed using the continuous model described. Samples were amplified following the manufacturer's recommended protocols (except that PCR volume was halved to be 25 μL) with Applied Biosystems' Profiler Plus® multiplex (Life technologies, Carlsbad, CA) on an Applied Biosystem's 9700 (silver block) thermal cycler. Amplified DNA was separated on an Applied Biosystem's 3130xl capillary electrophoresis instrument. Resulting DNA profiles were analysed using Applied Biosystem's GeneMapper ID v3.2.1. EPGs were analysed using the continuous model with 50 000 accepts (of which the first 10 000 were burn-in). The samples in this study resulted in 39, 274, 207, and 50 comparisons to single source, two, three and four person mixed profiles, respectively. LR_C produced by the model were compared to the human interpretation for the same result (Fig. 6). The hypotheses considered are:

H_1: the DNA came from POI and unknown people up to the number of contributors.

H_2: the DNA came from all unknown people.

Human interpretations were sorted into three categories; "not excluded", "inconclusive" or "excluded".

Inspection of the graph shows a broad alignment of human and model based interpretation except that on average human interpretations were more conservative.

References

[1] T. Clayton, J.P. Whitaker, R.L. Sparkes, P. Gill, Analysis and interpretation of mixed forensic stains using DNA STR profiling, Forensic Sci Int 91 (1998) 55–70.
[2] M.W. Perlin, M.M. Legler, C.E. Spencer, J.L. Smith, W.P. Allan, J.L. Belrose, et al., Validating TrueAllele® DNA mixture interpretation, J Forensic Sci 56 (2011) 1430–1447.
[3] I.W. Evett, P.D. Gill, J.A. Lambert, Taking account of peak areas when interpreting mixed DNA profiles, J Forensic Sci 43 (1) (1998) 62–69.
[4] J.S. Buckleton, J.M. Curran, I.W. Evett, L.A. Foreman, S. Pope, C.M. Triggs, inventors; International Patent "Statistical Simulations in Mixtures" WO 03/0560352003.
[5] J.S. Buckleton, P.D. Gill, P17961 International Patent "Preferential degradation". 2003.
[6] I.W. Evett, L.A. Foreman, S. Pope, Use of Monte Carlo simulation to calculate likelihood ratios for mixed DNA profiles, 2001 (unpublished results).
[7] P. Gill, J.P. Whitaker, C. Flaxman, N. Brown, J.S. Buckleton, An investigation of the rigor of interpretation rules for STR's derived from less that 100 pg of DNA, Forensic Sci Int 112 (1) (2000) 17–40.
[8] J. Buckleton, P. Gill, Further comment on "Low copy number typing has yet to achieve "general acceptance" by Budowle, B. et al., 2009. Forensic Sci. Int. Genetics: Supplement Series 2, 551–552, Forensic Sci Int Genet 5 (1) (2011) 7–11.
[9] B. Budowle, R. Chakraborty, A. van Daal, Author's response to Gill P, Buckleton J. Commentary on: Budowle B, Onorato AJ, Callaghan TF, Della Manna A, Gross AM, Guerrieri RA, Luttman JC, McClure DL. Mixture interpretation: defining the relevant features for guidelines for the assessment of mixed DNA profiles in forensic casework, J Forensic Sci 55 (1) (2010) 265–268.
[10] B. Budowle, A. van Daal, Reply to comments by Buckleton and Gill on "Low copy number typing has yet to achieve 'general acceptance'" by Budowle, B. et al., 2009. Forensic Sci. Int.: Genet. Suppl. Series 2, 551–552, Forensic Sci Int Genet 5 (1) (2011) 12–14.
[11] D.J. Balding, J. Buckleton, Interpreting low template DNA profiles, Forensic Sci Int Genet 4 (1) (2009) 1–10.
[12] H. Haned, Forensim: an open-source initiative for the evaluation of statistical methods in forensic genetics, Forensic Sci Int Genet 5 (4) (2011) 265–268.
[13] K.E. Lohmueller, N. Rudin, Calculating the weight of evidence in low-template forensic DNA casework, J Forensic Sci 58 (1) (2013) S243–S249.
[14] B. Budowle, A.J. Onorato, T.F. Callagham, A.D. Manna, A.M. Gross, R.A. Guerreri, et al., Mixture interpretation: defining the relevant features for guidelines for the assessment of mixed DNA profiles in forensic casework, J Forensic Sci 54 (3) (2009).
[15] P. Gill, A. Kirkham, J. Curran, N. LoComatio, A software tool for the analysis of low copy number DNA profiles, Forensic Sci Int 166 (2–3) (2007) 128–138.
[16] J.S. Buckleton, C.M. Triggs, S.J. Walsh, Forensic DNA Evidence Interpretation, CRC Press, Boca Raton, Florida, 2004.
[17] J.-A. Bright, D. Taylor, J.M. Curran, J.S. Buckleton, Developing allelic and stutter peak height models for a continuous method of DNA interpretation, Forensic Sci Int Genet 7 (2) (2013) 296–304.
[18] J.-A. Bright, D. Taylor, J. Curran, J. Buckleton, Degradation of forensic DNA profiles, Aust J Forensic Sci (2013), http://dx.doi.org/10.1080/00450618.2013.772235.
[19] T. Tvedebrink, P.S. Eriksen, H.S. Mogensen, N. Morling, Amplification of DNA mixtures: missing data approach, Forensic Sci Int Genet Suppl Ser 1 (1) (2008) 664–666.
[20] T. Tvedebrink, P.S. Eriksen, H.S. Mogensen, N. Morling, Estimating the probability of allelic drop-out of STR alleles in forensic genetics, Forensic Sci Int Genet 3 (4) (2009) 222–226.
[21] C. Brookes, J.-A. Bright, S. Harbison, J. Buckleton, Characterising stutter in forensic STR multiplexes, Forensic Sci Int Genet 6 (1) (2012) 58–63.
[22] P.S. Walsh, N.J. Fildes, R. Reynolds, Sequence analysis and characterization of stutter products at the tetranucleotide repeat locus vWA, Nucleic Acids Res 24 (1996) 2807–2812.

[23] D.J. Balding, R.A. Nichols, DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands, Forensic Sci Int 64 (1994) 125–140.
[24] D.J. Balding, Estimating products in forensic identification using DNA profiles, J Am Stat Assoc 90 (431) (1995) 839–844.
[25] J.M. Curran, J.S. Buckleton, C.M. Triggs, B.S. Weir, Assessing uncertainty in DNA evidence caused by sampling effects, Sci Justice 42 (1) (2002) 29–37.
[26] J.M. Curran, J.S. Buckleton, An investigation into the performance of methods for adjusting for sampling uncertainty in DNA likelihood ratio calculations, Forensic Sci Int Genet 5 (5) (2011) 512–516.
[27] J.H. Ahrens, U. Dieter, Generating gamma variates by a modified rejection technique, Commun Aust Res Council 25 (1982) 47–54.
[28] J.H. Ahrens, U. Dieter, Computer methods for sampling from gamma, beta, poisson and binomial distributions, Computing 12 (1974) 223–246.
[29] D.J. Balding, likeLTD (likelihoods for low-template DNA profiles), 2013 Available from: https://sites.google.com/site/baldingstatisticalgenetics/software/likeltd-r-forensic-dna-r-code.
[30] H. Kelly, J.-A. Bright, J. Curran, J. Buckleton, The interpretation of low level DNA mixtures, Forensic Sci Int Genet 6 (2) (2012) 191–197.

2.6 – clarification

Point 1: Formal description of the LR:

I provide a formal description of the LR below (however, I will use O instead of Gc, which is consistent with later works):

We seek the likelihood ratio (across all loci):

$$LR = \frac{\Pr(\mathbf{O} \mid Hp)}{\Pr(\mathbf{O} \mid Hd)}$$

Let there be J different genotype sets (S) of N contributors that can be considered $\{\mathbf{S}_j : j = 1...J\}$, so that:

$$LR = \frac{\sum_{j=1}^{J} \Pr(\mathbf{O} \mid \mathbf{S}_j, Hp) \Pr(\mathbf{S}_j \mid Hp)}{\sum_{j=1}^{J} \Pr(\mathbf{O} \mid \mathbf{S}_j, Hd) \Pr(\mathbf{S}_j \mid Hd)}$$

Noting that once genotype sets are specified the probability of the observed data is no longer dependant on the hypothesis:

$$LR = \frac{\sum_{j=1}^{J} \Pr(\mathbf{O} \mid \mathbf{S}_j) \Pr(\mathbf{S}_j \mid Hp)}{\sum_{j=1}^{J} \Pr(\mathbf{O} \mid \mathbf{S}_j) \Pr(\mathbf{S}_j \mid Hd)}$$

Note that J can be very large. For each locus, where there are 'a' possible allele a contributor can possess $\frac{a(a+1)}{2}$ different genotypes (obtained by the number of pairwise comparisons between a elements plus a homozygous genotypes). An N person mixture at L loci will possess $\left[a\left(\frac{a+1}{2}\right)\right]^{LN}$ possible genotypes sets, so if we take a modern multiplex that possesses approximately 20 loci, each with approximately 15 alleles $J > 10^{124}$. Many of these will not contribute to either one or both of the sums in the LR because:

1. $\Pr(\mathbf{O} \mid \mathbf{S}_j) \approx 0$, if the probability of the observed data is so low given genotype set j that it is approximately 0 i.e. the genotype set requires so much peak height variability, drop-in, drop-out or other improbable DNA profile events.
2. $\Pr(\mathbf{S}_j \mid H) = 0$, if the proposition requires the contribution of DNA from an individual whose genotype is not represented in set j.

It may be useful to think of the sum across j in the LR to be across all genotype sets where:

$$\Pr(\mathbf{O} \mid \mathbf{S}_j) \Pr(\mathbf{S}_j \mid H) > 0$$

However, it needs to be realised that the number of non-zero elements that would apply to the numerator and denominator could (and usually would) be different due to the second condition

above being unique to each proposition. It therefore may be useful to think of J as the number of genotype sets for which $\Pr(\mathbf{O} \mid \mathbf{S_j}) > 0$, so that the sum is over the same number of genotype sets in numerator and denominator but may still possess some zero elements due to the second condition above.

Each genotype set $\mathbf{S_j}$ contains N elements (genotypes) $\mathbf{S_j} = \left\{ {}^{1}G, ..., {}^{N}G \right\}$, where a left superscript is used to denote a contributor position. Also note the shift from **S** (to denote a set) to G (to denote a genotype). When a genotype set is conditional on a proposition, it can be broken down by $\mathbf{S_j} \mid Hp = \left\{ \mathbf{S}_{u,j\mid Hp}, \mathbf{S}_k, \mathbf{S}_p \right\}$, where:

- $\mathbf{S_k}$ are the set of known contributors (under both Hp and Hd) to the sample
- $\mathbf{S_p}$ are the set of persons of interest that are contributors under Hp but not Hd and
- $\mathbf{S_{u,j|Hp}}$ are the set unknown individuals from the population so that $\left|\mathbf{S}_{u,j}\right| + \left|\mathbf{S}_k\right| + \left|\mathbf{S}_p\right| = N$. Note that this term still requires a reference to the genotype set j, as (unlike the known contributors or persons of interest) the genotypes of unknown individuals change with changing j.

Note that $\mathbf{S_p}$ does not apply to Hd (i.e $\mathbf{S_j} \mid Hd = \left\{ \mathbf{S}_{u,j\mid Hd}, \mathbf{S}_k \right\}$). Due to this, $\mathbf{S_{u,j}}$ is therefore also different under Hp and Hd. I identify this difference by specifying $\mathbf{S_{u,j|Hp}}$ and $\mathbf{S_{u,j|Hd}}$ for Hp and Hd respectively. Note that both $\mathbf{S_k}$ and $\mathbf{S_{u,j|Hp}}$ can be empty sets, however $\mathbf{S_p}$ and $\mathbf{S_{u,j|Hd}}$ must contain at least one element in a forensic evaluation. Also note that $\mathbf{S}_{u,j\mid Hd} = \left\{ \mathbf{S}_{u,j\mid Hp}, \mathbf{S}_p \right\}$. This gives LR:

$$LR = \frac{\sum_{j=1}^{J} \Pr(\mathbf{O} \mid \mathbf{S_j}) \Pr(\mathbf{S}_{u,j\mid Hp}, \mathbf{S}_k, \mathbf{S}_p \mid Hp)}{\sum_{j=1}^{J} \Pr(\mathbf{O} \mid \mathbf{S_j}) \Pr(\mathbf{S}_{u,j\mid Hd}, \mathbf{S}_k \mid Hd)}$$

If we then consider that we have the reference information for the individuals in $\mathbf{S_p}$ and $\mathbf{S_k}$ then:

$$LR = \frac{\sum_{j=1}^{J} \Pr(\mathbf{O} \mid \mathbf{S_j}) \Pr(\mathbf{S}_{u,j\mid Hp}, \mathbf{S}_k, \mathbf{S}_p \mid \mathbf{S}_k, \mathbf{S}_p, Hp) \Pr(\mathbf{S}_k, \mathbf{S}_p)}{\sum_{j=1}^{J} \Pr(\mathbf{O} \mid \mathbf{S_j}) \Pr(\mathbf{S}_{u,j\mid Hd}, \mathbf{S}_k \mid \mathbf{S}_k, \mathbf{S}_p, Hd) \Pr(\mathbf{S}_k, \mathbf{S}_p)}$$

The probability of the genotypes of individuals in $\mathbf{S_p}$ and $\mathbf{S_k}$ is independent of proposition and so cancels in the numerator and denominator of the LR. Additionally, under Hp individuals in $\mathbf{S_p}$ and $\mathbf{S_k}$ are known to contribute, so that $\Pr(\mathbf{S}_{u,j\mid Hp}, \mathbf{S}_k, \mathbf{S}_p \mid \mathbf{S}_k, \mathbf{S}_p, Hp) = \Pr(\mathbf{S}_{u,j\mid Hp} \mid \mathbf{S}_k, \mathbf{S}_p, Hp)$, as $\Pr(\mathbf{S}_k, \mathbf{S}_p \mid \mathbf{S}_k, \mathbf{S}_p, Hp) = 1$, giving LR:

$$LR = \frac{\sum_{j=1}^{J} \Pr(\mathbf{O} \mid \mathbf{S_j}) \Pr(\mathbf{S}_{u,j|Hp} \mid \mathbf{S}_k, \mathbf{S}_p, Hp)}{\sum_{j=1}^{J} \Pr(\mathbf{O} \mid \mathbf{S_j}) \Pr(\mathbf{S}_{u,j|Hd} \mid \mathbf{S}_k, \mathbf{S}_p, Hd)}$$

Point 2: Equation 8 on paper page 523

This equation should read (in line with the previous clarification point):

$$LR = \frac{\sum_{j=1}^{J} \Pr(\mathbf{O} \mid \mathbf{S_j}, \mathbf{M}) \Pr(\mathbf{S}_{u,j|Hp} \mid \mathbf{S}_k, \mathbf{S}_p, Hp)}{\sum_{j=1}^{J} \Pr(\mathbf{O} \mid \mathbf{S_j}, \mathbf{M}) \Pr(\mathbf{S}_{u,j|Hd} \mid \mathbf{S}_k, \mathbf{S}_p, Hd)}$$

Point 3: Appendix B

I provide a replacement to appendix B below (but have replace *E*, which was used to denote evidence, with ***O***, used to denote observed data in order to be more consistent with later works):

The *LR* seeks to calculate:

$$LR = \frac{\Pr(\mathbf{O} \mid H_1)}{\Pr(\mathbf{O} \mid H_2)}$$

where H_1 and H_2 specify two propositions, typically those of prosecution and defence. We can consider a number of nuisance variables required to evaluate the probabilities in the *LR*. The most commonly considered variable within forensic genetics is the genotype sets that the contributors could possess. These will specify which alleles are expected to be present or absent from the profile, but not their expected heights. Introducing genotype sets ($\mathbf{S_j}$) in the *LR* gives:

$$LR = \frac{\sum_{j} \Pr(\mathbf{O} \mid \mathbf{S_j}) \Pr(\mathbf{S_j} \mid H_1)}{\sum_{j} \Pr(\mathbf{O} \mid S_j) \Pr(\mathbf{S_j} \mid H_2)}$$

Note:

- The $\Pr(\mathbf{O} \mid \mathbf{S_j})$ does not depend on the hypothesis and so they are removed from the conditioned terms. This is because the hypothesis denotes genotypes sets, so both are not required.
- We have not discriminated between genotype sets in the numerator and denominator. This is because the probabilities $\Pr(\mathbf{O} \mid \mathbf{S_j})$ are independent of hypothesis.

We now consider parameters that are used to describe the peak height data in a DNA profile. They include:

- Template DNA amount for each contributor (*n*), which has prior $t_n \sim U[0,T]$ (where T represents the upper limit on template amount before a DNA profile will no longer be analysed and is termed a saturation level). Let **T** be the set of N template values.
- Degradation for each contributor, which has prior $d_n \sim U[0,D]$ (where *D* represents a level of degradation above which profiles will generally be considered too low quality and will not be analysed). Let **D** be the set of *N* degradation values.
- A PCR replicate efficiency term for each PCR replicate (*y*), which has prior $R_y \sim U[0,\infty]$ (note that in practise, if an analysis was carried out and a replicate amplification efficiency obtained beyond the approximate bounds [0.1,10] it would be considered that one of the replicates is likely to have been the subject of an amplification error and should not be included in the analysis). Let **R** be the set of *Y* replicate amplification efficiency values.
- An amplification efficiency term for each locus (*l*), which has prior $A^l \sim LN(0,\xi^2\sigma^2)$ (where $\xi = \ln(10)$ is used to transform between logs in base 10 and base *e* and σ^2 is determined by laboratory calibration). Let **A** be the set of L locus amplification efficiency values.
- A peak height variability parameter for each fluorescence type (*i*), which has prior $c^i \sim \Gamma(\alpha^i,\beta^i)$ (which is determined by laboratory calibration). Let **C** be the set of *I* peak height variability values.

Let $\mathbf{M} = \{\mathbf{T},\mathbf{A},\mathbf{R},\mathbf{D},\mathbf{C}\}$, which we term mass parameters. Including these nuisance parameters in the *LR* gives:

$$LR = \frac{\int \sum_j p(\mathbf{O} \mid \mathbf{S_j}, \boldsymbol{M}) \Pr(\mathbf{S_j} \mid, H_1) \Pr(\boldsymbol{M}) d\boldsymbol{M}}{\int \sum_j p(\mathbf{O} \mid \mathbf{S_j}, \boldsymbol{M}) \Pr(\mathbf{S_j} \mid, H_2) \Pr(\boldsymbol{M}) d\boldsymbol{M}}$$

Which is the form of the *LR* given in the body of the paper in equation (2). This integral, if **M** was expanded out into individual parameters, is high dimensionality a multidimensional integral with $2N + L + Y + I$ dimensions.

Due to the high dimensionality of the integration required, numerical Monte Carlo integration is infeasible. We instead use Markov Chain Monte Carlo (MCMC). MCMC sets up a posterior distribution as its limiting distribution. We use the Metropolis-Hastings algorithm so that after sufficient run time the Markov chains are sampling from the joint posterior distribution.

The integral could be evaluated separately for the numerator and denominator of the *LR*. It is expected that the posterior distribution for parameters within **M** would be similar between the integrals. The main difference would be that different individuals are specified in the propositions and so the prior probabilities for genotype sets will differ i.e. typically the prosecution proposition specifies additional contributors of DNA as known individuals and so it is expected that a larger number of genotype set prior probabilities would be zero (those that did not contain the genotypes of the specified individuals).

To do this would mean that for every comparison to a person of interest in a case, a separate integration would be required, which has associated time and computer resource costs. To overcome this, we use MCMC to evaluate the integral that does not take the two competing hypotheses into account. We do so by considering only genotype sets that satisfy conditions that the profile has originated from *N* contributors, the genotypes of some of which are fixed (those agreed to be contributors under both propositions). We therefore evaluate the integral:

$$\int \sum_j p(\mathbf{O} \mid \mathbf{S_j}, \boldsymbol{M}) \Pr(\boldsymbol{M}) d\boldsymbol{M}$$

The process for a single MCMC iteration is:

- Draw values for parameters within **M** by random walk
- Randomly choose a genotype set at one randomly chosen locus (leaving genotype set at all other loci unchanged) by choosing from any of the available genotype sets with equal probability, i.e. choose *l*, where $l \in [1, L]$. Let there be J^l genotypes at locus *l* then choose j^l so that $j^l \in [1, J^l]$
- Evaluate $p(\mathbf{O} \mid \mathbf{S_j}, \boldsymbol{M}) \Pr(\boldsymbol{M})$
- Accept or reject proposed parameters by Metropolis-Hasting algorithm

Note that only one genotype set is proposed within an iteration of the MCMC algorithm (as opposed to calculate the sum across all genotype sets at each iteration). Doing so decreases the acceptance rate, however this is offset by being able to complete an iteration with much less calculation. We found that the time increase due to a lower acceptance rate is less than the speedup due to quicker calculation time as there can be multiple genotype sets that have similarly high posterior probabilities.

Genotype sets can take only discrete, unordered values. The mean posterior probability for genotype set *j* can be determine by residence time of genotype set *j* as during the MCMC. The residence time of S_j in the MCMC will be directly related to its probability as this is what the Metropolis Hastings acceptance/rejection criteria are based on. In other words, residence time for genotype set *j*, r_j:

$$r_j \propto \int \sum_j p(\mathbf{O} \mid \mathbf{S_j}, \boldsymbol{M}) \Pr(\boldsymbol{M}) d\boldsymbol{M}$$

Note that by choosing genotype sets uniformly across all available sets we use proposed distribution *q*(*x*) that has been weighted compared to the target distribution $\pi(x)$ for a genotype set *j* at locus *l* by:

$$w = \frac{q(x)}{\pi(x)} = \frac{\left(\frac{1}{J^l}\right)}{\Pr(\mathbf{S_j^l} \mid H)}$$

for each $\mathbf{S_j^l}$ within the model. The choice to do this has the advantage that allele frequencies, and hence a population, does not need to be specified in the MCMC. The population, or indeed multiple populations, of interest can be chosen at a later time when an *LR* is required.

We correct for the bias introduced into the by multiplying each of the *J* posterior elements by the weight above within the *LR* to recover:

$$LR = \frac{\left(\frac{1}{J^l}\right)\int \sum_j p(\mathbf{O} \mid \mathbf{S_j}, \boldsymbol{M}) \Pr(\mathbf{S_j} \mid, H_1) \Pr(\boldsymbol{M}) d\boldsymbol{M}}{\left(\frac{1}{J^l}\right)\int \sum_j p(\mathbf{O} \mid \mathbf{S_j}, \boldsymbol{M}) \Pr(\mathbf{S_j} \mid, H_2) \Pr(\boldsymbol{M}) d\boldsymbol{M}}$$

$$= \frac{\sum_{j=1}^{J} \Pr(\mathbf{S_j} \mid Hp) \int p(\mathbf{O} \mid \mathbf{S_j}, \mathbf{M}) \Pr(\mathbf{M}) d\mathbf{M}}{\sum_{j=1}^{J} \Pr(\mathbf{S_j} \mid Hd) \int p(\mathbf{O} \mid \mathbf{S_j}, \mathbf{M}) \Pr(\mathbf{M}) d\mathbf{M}}$$

Chapter 3: The likelihood ratio

Even when manual systems of DNA profile interpretation were in use, the dominant form of reporting DNA profiling results in Australia was the LR. In fact, all but one laboratory in the years prior to 2012 (when STRmix™ was adopted) used this form of evaluation. Within the framework of reporting the strength of DNA evidence in LR form there are numerous topics of discussion that range from almost philosophical, to biological to outright mathematical. The publications within this chapter touch on a few of these topics, namely; the formation of propositions under which the findings will be considered, the numerical calculation of the LR for complex mixtures, the consideration of relatives of a person of interest and the sensitivity of the LR to prior distributions of parameters within the biological and statistical models.

There were different motivations that lead to many of these works and they are given throughout the chapter as it progresses.

3.1: The formulation of propositions

The LR requires the formation of two competing scenarios (called hypotheses or propositions within Forensic Biology circles) for the evidence. In the early years of DNA profiling, proposition formation was generally a relatively straightforward task as the DNA profiles that were considered for a numerical evaluation were restricted to such high quality and low complexity that the appropriate propositions to use were obvious. For example (and putting aside single source profiles, where the choice of propositions is completely obvious at sub-source level) a typical scenario where a DNA profile of sufficient quality was obtained would be an intimate swab from the victim of an alleged rape. The mixture obtained could be explained by the victim and suspect and was intense enough that it could be safely assumed all data was present (i.e. no dropout could have occurred). The propositions would then be:

1) The DNA came from the victim and suspect
2) The DNA came from the victim and an unknown male

The advent of STRmix™ meant that many more complex profiles could be evaluated, which brought with it the question of what propositions were appropriate. For example, imagine the same scenario as previously described however there is a single additional weak allele in the profile that indicated a third contributor. Further imagine that the victim's boyfriend (with whom she is sexually active) possesses this allele (along with approximately 30% of the population). Should the propositions now be:

1) The DNA came from the victim, boyfriend and suspect
2) The DNA came from the victim, boyfriend and an unknown male

It may be believed that there is insufficient information in this third, weak minor component to assume the boyfriend. Extend the scenario to one where four weak peaks were present that matched the boyfriend, or eight, or 16, etc Common questions that arise in proposition setting are: '*At what arbitrary point is enough to assume?*" and '*should the profile itself even be used to determine propositions?*' Leading philosophy on the topic suggests that propositions cannot be findings-lead. Instead, propositions should be set on case circumstances. However, if a decision has been made, based on case circumstances alone, to assume the presence of the boyfriend, and a DNA profile is received that shows no sign of a contribution of DNA by him, can the propositions be changed at this point?

Such vacillations made it appear as though the forensic biology community was still destined to be plagued by subjective and inconsistent choices of propositions, some of which would no doubt be based on largely arbitrary threshold-based decision.

An increasingly paced influx of such questions lead to the work presented in this section. This paper outlines the existing 'rules' of evidence evaluation and builds on them in light of the new-found ability to evaluate complex DNA profiles.

Manuscript: Helping formulate propositions in forensic DNA analysis. J Buckleton, JA Bright, D Taylor, I Evett, T Hicks, G Jackson, JM Curran. (2014) Science & Justice 54 (4), 258-261 – *Cited 7 times*

Statement of novelty: The work builds on previous philosophical recommendations on proposition development for use in LRs. In particular, it provides insights on how to formulate propositions in complex situations

My contribution: I was a roughly equal co-contributor to the theorising and writing of the paper.

Research Design / Data Collection / Writing and Editing = 15%/ NA / 15%

Additional comments:

Science and Justice 54 (2014) 258–261

Contents lists available at ScienceDirect

Science and Justice

journal homepage: www.elsevier.com/locate/scijus

Helping formulate propositions in forensic DNA analysis

John Buckleton [a,*], Jo-Anne Bright [a,b], Duncan Taylor [c], Ian Evett [d], Tacha Hicks [e], Graham Jackson [f,g], James M. Curran [b]

[a] ESR Ltd., Private Bag 92021 Auckland, New Zealand
[b] Department of Statistics, University of Auckland, Private Bag 92019, Auckland, New Zealand
[c] Forensic Science South Australia, 21 Divett Place, SA 5000, Australia
[d] Principal Forensic Services Ltd., London, UK
[e] University of Lausanne, Institut de police scientifique et Fondation pour la formation continue universitaire lausannoise, Lausanne-Dorigny, Switzerland
[f] 'Advance Forensic Science', St. Andrews, Fife, UK
[g] School of Contemporary Sciences, University of Abertay Dundee, Bell Street, Dundee DD1 1HG, UK

ARTICLE INFO

Article history:
Received 20 May 2013
Received in revised form 23 December 2013
Accepted 18 February 2014

Keywords:
Forensic DNA interpretation
Mixtures
Propositions
Investigation
Evaluation
Role

ABSTRACT

The Bayesian paradigm is the preferred approach to evidence interpretation. It requires the evaluation of the probability of the evidence under at least two propositions. The value of the findings (i.e., our *LR*) will depend on these propositions and the case information, so it is crucial to identify which propositions are useful for the case at hand. Previously, a number of principles have been advanced and largely accepted for the evaluation of evidence. In the evaluation of traces involving DNA mixtures there may be more than two propositions possible. We apply these principles to some exemplar situations. We also show that in some cases, when there are no clear propositions or no defendant, a forensic scientist may be able to generate explanations to account for observations. In that case, the scientist plays a role of investigator, rather than evaluator. We believe that it is helpful for the scientist to distinguish those two roles.

1. Introduction

Jackson et al. [1] have shown that a forensic scientist may have two roles: investigator or evaluator. For evaluation, a number of principles have been advanced and largely accepted [2–5]. More recently Berger et al. [6] offered the three principles for the interpretation of forensic evidence given below which comprise an extension of the earlier work by Evett and Weir [2]:

1) To form an evaluative opinion from a set of observations, it is necessary for the forensic scientist to consider those observations in the light of propositions and forensically relevant case information. The propositions should represent the positions of the different participants in the legal process. In a criminal trial, the propositions will represent the positions of prosecution and defence, respectively.
2) It is necessary for the scientist to consider the *probability of the observations* given each of the stated propositions. Not only is it not appropriate for the scientist to consider the *probability of the proposition* given the observations, there is a danger that in doing so the jury will be misled.
3) The ratio of the probability of the observations given the prosecution proposition to the probability of the observations given the defence proposition, which is known as the *likelihood ratio*, provides the most appropriate foundation for assisting the court in establishing the weight of the findings.

It is crucial that close attention is given to the formulation of propositions.

We are motivated in writing this paper by a complex murder case which we amend to conceal the actual origin but retain the salient features. We consider the murder of 4 related people. A bloodstain was recovered associated with the accused that could be explained as a mixture of all 4 deceased. However 27 people from the pedigree were sampled by the authorities and given as reference samples. Taken individually, 7 of these cannot be excluded from the mixture. We are aware of a policy [7–11] that takes each of the 27 reference samples individually and forms an *LR* from the hypotheses:

H1. The DNA comes from person m and 3 unknown persons.

H2. The DNA comes from 4 unknown persons.

* Corresponding author at: ESR, Private Bag 92021, Auckland 1142, New Zealand. Tel.: +6498153904.
E-mail address: john.buckleton@esr.cri.nz (J. Buckleton).

http://dx.doi.org/10.1016/j.scijus.2014.02.007
1355-0306/

This produced 27 likelihood ratios of which 7 were greater than 1. This was not unexpected given the number of potentially related people contributing DNA to the evidence sample.

Several of the 7 non-excluded people were overseas and not plausible donors to the mixed bloodstain. We would suggest that they should not have ever been considered.

Our goal is to discuss proposition formation in mixed DNA casework, and to differentiate the processes we follow depending on our roles (i.e., investigator or evaluator). Although we mainly focus on DNA and source level propositions, many of the principles involved, we believe are general to all transfer evidence.

2. Importance of the framework of circumstances and propositions

Before discussing propositions themselves we would like to underline the importance of both case information and propositions: propositions emanate from the case information given by both parties. They summarise both points of views and reflect the issues facing the court. On this topic (i.e., case information), we would like to underline that we do not consider information such as prior conviction, motive, presence of other types of evidence, or a confession as relevant forensic information: these are only the concern of the Court and forensic scientists should not be given this type of information. Forensically relevant case circumstances provide the information that will help in assigning probabilities for the observations; for example the delay between the offence and the seizure of the shoes (in a footwear case), or if a person suspected of having broken a window has activities that may lead to the transfer of broken glass, or in a DNA case information that will help formulate the appropriate alternative, determine the number(s) of contributors, and select the relevant population.

Returning to the case that motivated this paper we would suggest that the prosecution may sensibly allege that the blood is a mixture of the 4 deceased. The remaining 23 reference samples from the pedigree, whether excluded or not, are irrelevant and this is true whether or not they were in the country at the relevant time. The defence cannot reasonably concede that the blood is from any of the deceased. Nor is it particularly credible that the blood is a mixture of other relatives of the deceased either in or out of country. The case circumstances, therefore, virtually dictate the form of the propositions in this example.

Next we discuss the concept of the hierarchy of propositions [12,13]. Propositions were initially classified into three levels: offence, activity, or source. The top of the hierarchy is taken to be the offence level where the issue is one of guilt or innocence. An example of this could be 'the suspect raped the victim' or 'Mr. G murdered the victim' or 'Mr. S stole the car'. It is often held that this level of proposition is for the courts to consider and above the level at which a forensic scientist would usually operate. This is not strictly correct. Scientists can assess the value of their findings given offence propositions, as long as they use specialised knowledge and add value. Of course, scientists will not evaluate the probability of the propositions themselves, nor should they for any propositions, even sub-source.

An example of activity level propositions would be: 'the suspect had intercourse with the victim versus they only had social contact'. This differs from the offence level in that it talks about an activity (intercourse) without taking into account issues of the intent of the perpetrator (rape) or the consent of the complainant. The case circumstances would give information on the alleged activities, time lapse, and what is meant by social contact.

The first level in the hierarchy is taken to be the source level. At this level we consider questions of the type: 'did this semen come from Mr. A or did it come from an unknown person?' Considerations at this level do not directly relate to activity, in this example 'intercourse', which would involve issues such as when the specimen was taken, drainage, and contamination.

In the light of the sensitivity of DNA methods that now allows analysis of samples that cannot be detected with the naked eye (i.e., trace DNA), it has become necessary to add another level below the source level. This has been termed "sub-source" and has arisen because it is not always possible to infer from what body fluid the DNA has come [14]. For instance when considering the source level proposition 'the semen came from the suspect', the equivalent sub-source, or sub-level 1, proposition would be 'the DNA came from the suspect' and would not necessarily imply that the DNA came from semen. It could, alternatively, have come from saliva or epithelial cells.

The further down the hierarchy the scientist operates the more the responsibility for interpreting the evidence is transferred to the court or to other experts. It is therefore important that, if the assessment of the results demands forensic knowledge (e.g., factors such as transfer, persistence, presence of material for reasons unconnected to the alleged offence), forensic scientists help the Court to the best of their abilities and that they explain clearly what the results mean. Consequently, we would expect that forensic scientists help assess activity level propositions, if the above factors have a significant impact on the understanding of the alleged activities and require expert knowledge.

This led to the formalisation of an additional requirement for interpretation. Whilst not strictly a principle as given above we number it in the same sequence:

4) Due attention must be paid to the position in the hierarchy of propositions that can be considered. This information must be effectively conveyed to the court to avoid the risk that an evaluation at one level is translated uncritically and without modification to evaluation at a higher level.

We cannot over-emphasise the importance of this. A DNA match may inform decisions about the source of the DNA, but decisions about an activity, say sexual intercourse versus social contacts, involve additional considerations beyond the DNA profile.

3. Mixed DNA profiles

Consider a case where the laboratory receives a trace (labelled 'Forensic specimen'), as well as DNA samples from the victim and the person of interest. No case information is given. The profile derived from the trace shows 3 or 4 alleles for several loci. From these results we can infer that the profile is a mixture and that it comes from at least two persons.

Without any information on the trace or the circumstances, the profile may be a mixture of the victim, V, and the person of interest, P. Plausible hypotheses could be:

H1. The DNA came from V and P.

H2. The DNA came from V and an unknown person U_1.

H3. The DNA came from an unknown person U_2 and P.

H4. The DNA came from two unknown persons U_1 and U_2.

Are there any principles which may be applied to guide the choice of which pair of propositions should be used?

The key point here is to identify the issue that the Court wants to solve and to do so, one needs case information from both points of view. It is likely that the prosecution have formed their hypothesis and this is often known to the forensic scientist. It is unlikely that the defence will have formed their hypothesis and there is no requirement for them to do so. Under these circumstances we would offer the following as guidelines:

5) The prosecution proposition should be set to align with the prosecution's allegation and the case information.
6) There is no requirement for the defence to set or disclose their proposition. The forensic scientist should select a reasonable proposition consistent with defence's view (and again case information). If

this proves to be a poor choice subsequently, the findings should be re-assessed.

Recommendations 5 and 6 are consistent with the earlier recommendations of the DNA Commission of the ISFG (their recommendation 5 and appendix C) [4] and with the approaches outlined in Gill and Haned [15]. However we would suggest that they are not universally applied [7–11,16] and it may be necessary to reemphasise them. It may also be necessary to mention again the importance of case information (given by both parties) for helping to determine the number of contributors as this will impact on how the propositions are formulated. The sub-source proposition most beneficial to the defendant is likely to be the one that concedes the maximum number of known contributors consistent with the mixture not including the accused, minimising the difference in the number of known contributors between the prosecution and defence propositions. The addition of unknown contributors to the defence proposition is most unlikely to increase the probability of the findings given this proposition unless it substantially improves the fit to the evidence.

If we apply these principles and guidelines to the hypothetical two person mixture described above, then we need to consider the circumstances of the case. Imagine that the sample is a vaginal swab taken from the complainant, C. Then, it seems a reasonable assumption that the DNA components matching C have indeed come from her. That assumption should be stated in the scientist's report and is rightfully open to challenge down the line. Therefore the propositions 'The DNA came from C and P' and 'The DNA came from C and an unknown person' are perfectly reasonable.

Next, imagine that the DNA sample was taken from a shirt belonging to the person of interest, P. Again it seems a reasonable assumption that the DNA components matching P have indeed come from the person of interest. Therefore the propositions 'The DNA came from C and P' and the DNA came from an unknown person (U_1) and P' are perfectly reasonable.

Last, let us consider that the sample is from an object, for example a cloth, neither associated with the complainant nor the person of interest. If we apply guideline 5 then it seems likely the prosecution will assert that the DNA is from C and P. However, in considering guideline 6, exoneration could follow if any of the propositions:

H2. The DNA came from C and an unknown person U_1.

H3. The DNA came from an unknown person U_2 and P.

H4. The DNA came from two unknown persons U_1 and U_2 were true.

The defence therefore have a right to any of H_2…H_4 if the case information suggests that this is sensible. Gill and Haned [15] suggest exploring all these options and we concur.

Papers, such as this one, often emphasise consultation with the prosecution and defence. This is a lofty ideal that is seldom achieved in practice. It assumes both willingness and an ability on the part of all parties to consult in a constructive manner. In our experience the issues discussed in this paper are not universally understood by forensic scientists and seldom considered at all by lawyers. The reality is that the scientist is often left to form the propositions from the case information with little support from either prosecution or defence.

4. Complex situations: investigation or evaluation?

We are aware that the scientist may be presented with what amounts to a trawl. A profile is developed that may be from *N* individuals. *M* persons of interest are offered with the suggestion that they may have been involved in some combination. This is the situation which motivated this paper, discussed in the introduction. However it could also apply to any mixture tested against a database.

A very useful differentiation is made regarding the role of the forensic examiner at different parts of the criminal justice process. This draws a distinction between the investigative and evaluative stages [1,17]. At the investigative stage a scientist may be asked to offer an opinion to advance an inquiry. At this point a lot more liberty may be afforded by the scientist and a rigorous application of the principles of interpretation is not required although of course sensible scientific statements should be adhered to. The evaluative stage describes the situation when a scientist is required to assist a court of law by providing an evaluative opinion with regard to the assistance that a set of scientific observations might provide in addressing matters that the court is deliberating. Then the scientist's evaluation should be governed by a set of logical principles.

To help address investigative issues, an examiner generally offers opinions in the form of explanations or, if prior information and expert knowledge is taken into account, in the form of posterior probabilities for explanations.

To help address evaluative issues, an examiner would address a pair of clearly stated propositions and consider the probability of the findings given each of the propositions. The opinion expressed would reflect the ratio of those two probabilities (the likelihood ratio).

There are limitations to both investigative and evaluative opinions and these should be made clear at the outset. Of interest in this circumstance, likelihood ratios require, firstly, consideration of propositions that are based on the case circumstances and the competing allegations and, secondly, reliable and valid assignment of probabilities for the observations.

The very complex situations described above sit much more easily at the investigative phase and more awkwardly at the evaluative phase.

If there are *M* persons of interest, to a *N* contributor profile, then there are many pairs of hypotheses that could be considered. For complex profiles with many persons of interest the number of proposition pairs could number in the hundreds or thousands. This is clearly too many for an exhaustive exploration of likelihood ratios. A strategy in current use and described above, tries each of the *M* persons of interest in turn in H_1 with the remaining *N*-1 contributors as unknown. H_2 is set as *N* unknown contributors. Since there are *M* persons of interest *M LR*s will be produced and plausibly reported. We will term this the search strategy but we suggest that it sits more comfortably as part of an investigative phase.

When faced with very complex situations we have often found it useful to attempt to remind ourselves of the principles of evidence interpretation before even trying to apply them. One principle is that the addition of any relevant and correct information, on average, improves the power of the *LR* to help differentiate the two hypotheses.

We suggest that information known, or reasonably assumed, to be correct should be included in proposition setting. This will include any situation where someone's DNA can be reasonably thought to be present on an item, such as an intimate swab, item of clothing, owned object or object that has known to be handled by an individual.

The question then remains as to what to do when there are multiple individuals for comparison to a profile when no-one can be reliably assumed to have contributed DNA to the sample. Our preference is to ignore the uncertain information that the other persons of interest may be contributors and we believe that no choice is left other than allow the search strategy. This is perfectly acceptable at the investigative phase. The reporting of propositions and *LR*s in this manner is also sustainable at the evaluative stage in that it is an honest report of the propositions and consequent *LR*. However we would advocate an additional check if *LR*s greater than 1 are produced for two or more individuals. We point out that it is possible to produce *LR*s greater than 1 for more than *N* persons of interest for an *N* person mixture. Such a situation risks ridicule without careful explanation by the scientist. In addition some combinations of the individuals with *LR*s greater than 1 may be incompatible. We suggest that a report could indicate:

1. Which combinations were compatible with the mixture and whether additional contributors were needed to explain the mixture, and

2. Report that some combinations of persons of interest may be incompatible but that checking is impractical, and
3. Indicate that the list of combinations is not meant to be exhaustive.

The alignment of this information with the investigative stage and poor alignment with the evaluative stage would, we hope, be apparent from this information. Further the report could

4. Suggest in this investigative report, that information is needed to properly inform the propositions.

5. Conclusion

The principles and guidelines given above have proven useful in casework applications in guiding the selection of propositions. Principles 1 through 6 have been published and are largely accepted but have not been applied universally. We would suggest that these principles should be incorporated into all casework and that an essential first phase of all interpretation is a thoughtful formation of relevant propositions taking into consideration case information (both from prosecution and defence when possible).

One should also distinguish the evaluation and investigation phase (for example when there are multiple possible suspects as illustrated above). In the former, scientists assess their findings given propositions and the framework of circumstances, in the latter one can explore multiple pairs of propositions and suggest explanations for the findings. Here, however, as in all investigations, these are only leads and as such can be misleading. When case information will have been gathered from the defence (perspective), then these findings will need to be formally evaluated.

Acknowledgements

This work was supported in part by grant 2011-DN-BX-K541 from the US National Institute of Justice. Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the U.S. Department of Justice.

References

[1] G. Jackson, S. Jones, G. Booth, C. Champod, I.W. Evett, The nature of forensic science opinion — a possible framework to guide thinking and practice in investigations and in court proceedings, Sci. Justice 46 (2006) 33–44.
[2] I.W. Evett, B.S. Weir, Interpreting DNA Evidence — Statistical Genetics for Forensic Scientists, Sinauer Associates, Inc., Sunderland, 1998.
[3] I.W. Evett, G. Jackson, J.A. Lambert, S. McCrossan, The impact of the principles of evidence interpretation on the structure and content of statements, Sci. Justice 40 (2000) 233–239.
[4] P. Gill, C.H. Brenner, J.S. Buckleton, A. Carracedo, M. Krawczak, W.R. Mayr, N. Morling, M. Prinz, P.M. Schneider, B.S. Weir, DNA commission of the International Society of Forensic Genetics: recommendations on the interpretation of mixtures, Forensic Sci. Int. 160 (2006) 90–101.
[5] P. Gill, L. Gusmão, H. Haned, W.R. Mayr, N. Morling, W. Parson, L. Prieto, M. Prinz, H. Schneider, P.M. Schneider, B.S. Weir, DNA commission of the International Society of Forensic Genetics: recommendations on the evaluation of STR typing results that may include drop-out and/or drop-in using probabilistic methods, Forensic Sci. Int. Genet. 6 (2012) 679–688.
[6] C.E.H. Berger, J. Buckleton, C. Champod, I.W. Evett, G. Jackson, Evidence evaluation: a response to the court of appeal judgment in R v T, Sci. Justice 51 (2011) 43–49.
[7] M.W. Perlin, The Blairsville Slaying and the Dawn of DNA Computing, in: A. Niapas (Ed.), Death needs answers: the cold-blooded murder of Dr John Yelenic, Grelin Press, New Kensington, 2013.
[8] M.W. Perlin, J. Galloway, Computer DNA evidence interpretation in the Real IRA Massereene terrorist attack, Evid. Technol. Mag. 10 (2012) 20–23.
[9] M.W. Perlin, Easy reporting of hard DNA: computer comfort in the courtroom, Forensic Mag. 9 (2012) 32–37.
[10] M.W. Perlin, DNA Identification Science, in: C.H. Wecht (Ed.), Forensic Sciences, LexisNexis Matthew Bender, Albany, 2012.
[11] M.W. Perlin, M.M. Legler, C.E. Spencer, J.L. Smith, W.P. Allan, J.L. Belrose, B.W. Duceman, Validating TrueAllele® DNA mixture interpretation, J. Forensic Sci. 56 (2011) 1430–1447.
[12] R. Cook, I.W. Evett, G. Jackson, P.J. Jones, J.A. Lambert, A hierarchy of propositions: deciding which level to address in casework, Sci. Justice 38 (1998) 231–240.
[13] I.W. Evett, G. Jackson, J.A. Lambert, More on the hierarchy of propositions: exploring the distinction between explanations and propositions, Sci. Justice 40 (2000) 3–10.
[14] I.W. Evett, P.D. Gill, G. Jackson, J. Whitaker, C. Champod, Interpreting small quantities of DNA: the hierarchy of propositions and the use of Bayesian networks, J. Forensic Sci. 47 (2002) 520–530.
[15] P. Gill, H. Haned, A new methodological framework to interpret complex DNA profiles using likelihood ratios, Forensic Sci. Int. Genet. 7 (2013) 251–263.
[16] M.W. Perlin, A. Sinelnikov, An information gap in DNA evidence interpretation, PLoS ONE 4 (2009) e8327.
[17] G. Jackson, Understanding Forensic Science Opinions, in: J. Fraser, R. Williams (Eds.), Handbook of Forensic Science, Willan Publishing, Cullompton, 2009.

3.2: A new level in the hierarchy of propositions brought about by continuous DNA profile interpretation

It became apparent from the work on proposition setting that there was something missing from the current picture. This was magnified as STRmix™ was being programming and references were being compared to complex mixtures (particularly when there was some support for them being any one of multiple contributors in the mixture). The logic went like this:

1) A reference profile is to be compared to a mixed evidence profile
2) There is *a priori* no reason to restrict comparisons to any particular component of the mixture
3) The more complex the mixture the more ways of comparing references to it and the more chance for them to 'match'
4) The LRs for these different ways of comparing the same evidence and reference profiles, using the same propositions, were different (except for the most contrived of circumstances)

This raises two problems. First, there are multiple LRs for the comparison of reference and evidence profiles, using the same propositions. Second, it is not clear which is the 'right' LR? Very early versions of STRmix™ simply reported the biggest of the multiple LRs obtained, however given the propositions were asking about an individual's potential contribution of DNA to the DNA profile as a whole (not specifying a component of it) and that in many instances the POI was excluded from being a contributor to other components of the mixture, there appeared to be information that was not being utilised.

The multiple mixture component situation is similar to the multi-testing problem associated with genetic tests for dependencies between loci. Simply put, the more tests that are carried out, the more likely it is that an association will be found, just by chance.

The effect being seen was the result of a previously unrecognised level in the hierarchy of propositions, one that sat below the lowest level then recognised. A description of the new, lower, hierarchical position and the mathematical process of moving from this level, up to the next (and in the process considering the comparison of the reference to all components of the DNA mixture) is given in the paper in this section of the thesis.

Manuscript: The ‘factor of two’ issue in mixed DNA profiles. D Taylor, JA Bright, J Buckleton. (2014) Journal of theoretical biology 363, 300-306 – *Cited 7 times*

Statement of novelty: This paper builds on the hierarchy of propositions concept and extends the theory to explain how it applies to sub-components of DNA profiles within a continuous DNA interpretation system.

My contribution: I was main author and main theorist on this work. I carried out the simulations that are used in the paper.

Research Design / Data Collection / Writing and Editing = 85%/ NA / 80%

Additional comments:

Journal of Theoretical Biology 363 (2014) 300–306

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

The 'factor of two' issue in mixed DNA profiles

Duncan Taylor [a,b,*], Jo-Anne Bright [c], John Buckleton [c]

[a] Forensic Science South Australia, 21 Divett Place, Adelaide, SA 5000, Australia
[b] School of Biological Sciences, Flinders University, GPO Box 2100, Adelaide, SA 5001, Australia
[c] Institute of Environmental Science and Research Limited, Private Bag 92021, Auckland 1142, New Zealand

ARTICLE INFO

Article history:
Received 5 May 2014
Received in revised form
7 August 2014
Accepted 11 August 2014
Available online 23 August 2014

Keywords:
Forensic DNA interpretation
Mixtures
Hierarchy of propositions

ABSTRACT

A commonly used idea in forensic fields is known as the 'hierarchy of propositions'. DNA analysts commonly report at the sub-source level in the hierarchy. This means that they simply comment on the probability of the evidence for the given propositions that consider contributors that lead to a DNA profile and not on the source of specific biological components, not the activity that led to the transfer or the offence that is reported to have occurred. However DNA analysts also commonly report at a level even lower than the sub-source level. In this 'sub–sub-source' level only reference comparisons to components of a mixture are reported. The difference between the sub-source level and sub–sub-source level is the difference between comparing an individual to a mixture as a whole, or comparing them to only one component of a mixture. This idea has been expressed in the past as the 'two trace' problem or the 'factor of two' problem. With the advent of expert systems that can provide a measure of weight of evidence in the form of a likelihood ratio (*LR*) for any mixture, resolvable or not, the distinction between these two levels becomes more important. In this paper we explore how the *LR* can be constructed to report correctly at the sub-source level, by taking contributor orders and genotype set orders into account. We include worked examples of the *LR* calculation to help explain this confusing issue.

1. Introduction

Previous literature has explored the classic two-trace transfer problem and explained that if multiple offenders are being considered then the propositions and therefore the resulting likelihood ratio (*LR*) need to consider this (Evett, 1987; Triggs and Buckleton, 2003; Triggs, 2004; Meester and Sjerps, 2003; Gittelson et al., 2012, 2013). The specific example considered is one where two stains and two offenders exist. If there are two relevant evidence DNA profiles suggesting two offenders, and one profile matches the only person of interest (POI) in the case then the *LR* has a factor of two in the denominator, which would not be there if only one DNA profile had been obtained. This has coined the term for this idea the 'factor of two' issue.

Given an observed crime stain ($\boldsymbol{O}$) an *LR* can be calculated considering two competing propositions, H_1 and H_2

$$LR = \frac{\Pr(\boldsymbol{O}|H_1)}{\Pr(\boldsymbol{O}|H_2)}$$

In the event that a single stain may be considered to comprise a clear major and minor component it may be thought of as two stains (Buckleton et al., 2004). However in general this is not possible and it should be thought of as one stain composed from two contributors. Owing to the fact that the components of a mixture can be fully or partially unresolved the work we present here differs from the classic two-trace problem, which deals with completely separate evidence electropherograms.

Regardless of whether the components of the mixture are resolvable we can consider the propositions:

H_1: the POI is one of the two persons in the mixture.
H_2: two unknown people are in the mixture.

We will term this proposition pair 1 (PP_1) and their use leads to LR_{PP1}. If the POI genotype (G_P) matches the major component (G_C^{Major}) of the observed stain and the genotype of the major contributor is unambiguous then the *LR* simplifies to

$$LR = \frac{1}{2\Pr(G_C^{Major})} \quad (1)$$

However, typically in forensic laboratories the *LR* that is reported is

$$LR = \frac{1}{\Pr(G_C^{Major})}$$

* Corresponding author at: Forensic Science South Australia, 21 Divett Place, Adelaide, SA 5000, Australia. Tel.: +61 8 8226 7700; fax: +61 8 8226 7777.
E-mail address: Duncan.Taylor@sa.gov.au (D. Taylor).

http://dx.doi.org/10.1016/j.jtbi.2014.08.021
0022-5193/

By not including the factor of two, the propositions being considered are

H_1: the POI is the donor of the major portion of DNA in the mixture.
H_2: an unknown person is the donor of the major portion of DNA in the mixture.

We will term this proposition pair 2 (PP_2, leading to LR_{PP2}). Often the use of PP_2 is not a conscious decision of the scientists, but rather an unrealised consequence of not including the factor of two in the calculation. We suggest the use of PP_2 is often not appropriate because it places responsibility on the jury to recognise the factor of two issue, which without specific attention drawn to that fact by the scientist, is unlikely to occur.

We have regularly lectured the use of *LR*s to audiences often unaccustomed to them. We would always use the PP_1 set. The audience often expresses surprise or even dismay at the factor of two. We diagnose this to arise from a mix of unfamiliarity with the literature on the factor of two, a history of use of PP_2, and concern at the apparently conflicting result. Even to the authors there appeared to be two different but equally valid approaches. For us, and we hope the readers, these views may be reconciled by considering the hierarchy of propositions (Cook et al., 1998a, 1998b; Berger et al., 2011; Evett et al., 2002, 2000; Jackson et al., 2006).

Propositions are classified into four levels: offence, activity, source, and sub-source. The top of the hierarchy is taken to be the offence level where the issue is one of guilt or innocence. An example of this could be 'Mr. Smith raped Ms. Doe' versus 'Mr. Smith had consensual sex with Ms. Doe.' The next level is taken to be the activity level. An example of this could be 'Mr. Smith had sexual intercourse with Ms. Doe' versus 'Mr. Smith had only social interaction with Ms. Doe.' The next level in the hierarchy is the source level. At this level we consider propositions of the type 'The semen came from Mr. Smith' versus 'The semen came from some unknown person'. The last level is termed sub-source. This arises when it is uncertain from what body fluid the DNA may have come. For instance, if one has recovered 'trace DNA', the sub-source propositions could be 'the DNA came from the suspect' versus 'the DNA came from someone else'.

To reconcile the different results using PP_1 and PP_2 we note that PP_1 is at sub-source level and PP_2 is something even lower. We term this sub–sub-source. We need to note that PP_2 propositions do not consider the entire DNA profile but only part of the DNA profile. Note that our intention is not to legitimise the intentional or unintentional use of the PP_2 propositions by identifying them as belonging to a 'sub–sub-source' level proposition category.

Buckleton et al. (2004, p. 255) had foreshadowed this distinction and made the same recommendation.

When a mixture is unresolved the POI's genotype appears as both contributor one and contributor two with equal weight, then $LR_{PP1} = LR_{PP2}$. This will be true for any system of interpretation that weights different genotype sets equally, such as the combined probability of inclusion (CPI) and the unconstrained combinatorial. It is generally recognised that systems that do not weight genotypes sets are wasteful of profile information (Perlin and Sinelnikov, 2009; Cowell et al., 2007; Evett et al., 1998; Gill et al., 1998). For many mixtures, depending on the level of resolution of their contributors, we will demonstrate that $LR_{PP1} < LR_{PP2}$. In the circumstance where $LR_{PP1} = LR_{PP2}$ it does not signify a shift in propositions being considered. All that has occurred is that the likelihood ratios considering either hypotheses are numerically equal.

Typically DNA profiling systems will target multiple regions, with currently available commercial multiplexes providing over 20 loci. The discrimination power of a DNA profile from such a system can be around 30 orders of magnitude. In such circumstances a difference of a factor of two makes no practical difference. We will show in this work that for more complex situations the difference can be larger and in one case we have worked on it was close to 20. Even at these levels this factor has little practical impact on the evidential strength of a large *LR*. For low level and partial profiles the relative importance of a change in *LR*s may be larger but it is unlikely the full factor will be needed since this occurs when there is a clear separation in heights between the contributors. Even so, a situation can be envisaged where the difference between LR_{PP1} and LR_{PP2} is from the hundreds to the tens, which may convey different evidential strengths to some people. The motivation for this work is not necessarily that it will have a large impact on justice outcomes, but rather to provide a framework for the calculation of an *LR* given propositions at different levels in the hierarchy. We demonstrate what factors will influence the size of the difference between LR_{PP1} and LR_{PP2} and when the effects will be at their greatest, so that this information can be used to generate approximations should a laboratory wish to apply them.

The complexity in generating *LR*s appropriately at the sub-source level and a general lack of understanding the difference between PP_1 and PP_2 within practising forensic scientists means LR_{PP1} is not widely considered if a major contributor can be resolved. Indeed even the mathematical solution we provide in this paper is likely to be beyond the reasonable expectation of 'by-hand' calculations and will require specifically designed computer software in order to implement.

There is a general move from binary systems to continuous systems in the interpretation of evidence. Binary systems assign either the value zero or one to the probability of the observed profile given a postulated set of genotypes, $\boldsymbol{S_j}$ (bolded to signify that it contains a set of multiple genotypes). Semi-continuous systems such as likeLTD (Balding, 2013; Balding and Buckleton, 2009), Lab Retriever (Lohmueller and Rudin, 2013) and LRmix (Haned, 2011) assign probabilities to posed genotype sets that are built up from a set of probabilities associated with drop-in, drop-out and stutter. Some semi-continuous systems have features that enable users to apply different probabilities to contributors in a mixture (such as the major and minor contributors) and hence introduce contributor order, which requires a careful consideration of proposition formation. Continuous DNA interpretation systems such as STRmix™ (Taylor et al., 2013) and True Allele™ (Perlin et al., 2011) can provide a likelihood (hereafter referred to as a weight) for the observed crime stain given proposed genotype sets that could explain it. Following the nomenclature of Taylor et al. (2013) we define these likelihoods as $p(\boldsymbol{O}|\boldsymbol{S_j}) \sim w_j$, and note that they definitely introduce contributor order into calculations. These likelihoods are seldom zero. This provides both the motivation and a means to account for the number of unknown contributors to the evidence profile. The motivation arises from the fact that a non-zero weight is a possible outcome for genotype sets with the POI aligned with either the contributor one genotype or the contributor two genotype.

We describe the methodology in this paper.

2. Findings

Deconvolution is a process which resolves an '*N*' person profile into '*J*' genotype sets ($\boldsymbol{S_{j\ j=1..J}}$) that each comprise '*N*' single person genotypes. Hence a certain set, say, $\boldsymbol{S_j}$ would have *N* genotypes G_1, G_2,…,G_N where the order of the genotypes assigns them to a specific contributor in the *N* person mixture.

For each genotype set, the likelihood $p(\boldsymbol{O}|\boldsymbol{S_j})$ is calculated. To present these densities in a manner that is more familiar with non-statisticians the probability densities for all genotype sets are normalised within each locus. This is true of the data we present in examples, when weights are referred to. In such a system, the

order of genotypes is important. Each element in the ordered set has associated with it a set of variables that collectively describe template at each allelic position. Some of these variables are contributor dependent and so pass a dependence on contributor order along to genotype sets.

There is a second ordering that we need to consider. If we consider that the set $\boldsymbol{S_j}$ specifies an order of genotypes that has been tested against an electropherogram then known and unknown contributors may be aligned with this set, also in different orders. We call the first ordering 'genotype order' and the second 'contributor order'. Contributor orders ($\mathcal{C}$) describe the arrangements that individuals can be compared to a list of genotype sets. For a single contributor profile there is a single position with which the POI can be compared and so only one contributor order exists, $\mathcal{C}_1$. For a two person profile there are two contributor orders $\mathcal{C}_1$ and $\mathcal{C}_2$ i.e. in contributor order one ($\mathcal{C}_1$) the POI is compared with the genotypes in position one and the unknown (U_1 for 'unknown 1') is compared to the genotypes in position two $\{\mathcal{C}_1 : (POI, U_1)\}$. The order is reversed in contributor order two $\{\mathcal{C}_2 : (U_1, \mathrm{POI})\}$. For three person profile there are six orders $\{\mathcal{C}_1 : (\mathrm{POI}, \mathrm{U}_1, \mathrm{U}_2), \mathcal{C}_2 : (\mathrm{POI}, \mathrm{U}_2, \mathrm{U}_1), \mathcal{C}_3 : (\mathrm{U}_1, \mathrm{POI}, \mathrm{U}_2), \mathcal{C}_4 : (\mathrm{U}_1, \mathrm{U}_2, \mathrm{POI}), \mathcal{C}_5 : (\mathrm{U}_2, \mathrm{POI}, \mathrm{U}_1), \mathcal{C}_6 : (\mathrm{U}_2, \mathrm{U}_1, \mathrm{POI})\}$, and for an '$N$' person profile there are $N!$ contributor orders $\{\mathcal{C}_i : i = 1, \ldots, N!\}$. We use the nomenclature of Taylor et al. (2013) by stating the continuous model LR as

$$LR_C = \frac{\sum_j w_j \Pr(\boldsymbol{S_j}|H_1)}{\sum_j w_j \Pr(\boldsymbol{S_j}|H_2)} \tag{2}$$

where $\Pr(\boldsymbol{S_j}|H_1)$ is the likelihood that N persons have the genotypes specified in $\boldsymbol{S_j}$ given H_1.

Following Taylor et al. (2013) we introduce the knowledge we have of the genotypes of the known contributor(s) and POI(s):

$$\Pr(\boldsymbol{S_j}|H_1) = \Pr(\boldsymbol{S_j}|\boldsymbol{S_K}, \boldsymbol{S_P}, H_1)\Pr(\boldsymbol{S_K}, \boldsymbol{S_P}|H_1)$$

where $\boldsymbol{S_P}$ is the genotype(s) of the POI (known contributor(s) under H_1 but not H_2), $\boldsymbol{S_K}$ is the genotype(s) of known contributors that are assumed to be present under both H_1 and H_2. We can then decompose the $\boldsymbol{S_j}$ into $\boldsymbol{S_P}$, $\boldsymbol{S_K}$ and the genotype(s) of unknown contributors, $\boldsymbol{S_U}$, so that

$$\Pr(\boldsymbol{S_j}|\boldsymbol{S_k}, \boldsymbol{S_p}, H_1) = \Pr(\boldsymbol{S_{U_j}}, \boldsymbol{S_p}, \boldsymbol{S_k}|\boldsymbol{S_k}, \boldsymbol{S_p}, H_1)$$

and

$$\Pr(\boldsymbol{S_{j'}}|\boldsymbol{S_k}, \boldsymbol{S_p}, H_2) = \Pr(\boldsymbol{S_{U_j}}, \boldsymbol{S_k}|\boldsymbol{S_k}, \boldsymbol{S_p}, H_2)$$

We define:

$$\Pr(\boldsymbol{S_K}|\boldsymbol{S_K}, H_1)\Pr(\boldsymbol{S_P}|\boldsymbol{S_P}, H_1) = \begin{cases} 1 & \{\boldsymbol{S_P} \subseteq \boldsymbol{S_j}\} \text{ and } \{\boldsymbol{S_K} \subseteq \boldsymbol{S_j}\} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

i.e. when the genotypes of known contributors and POIs are present in the genotype set the corresponding probability is assigned as one, and is not based on allele probabilities. If the genotypes of the known contributors or POIs are not in the genotype set then they are given probability of 0 for that proposition. This means that in Eq. (2)

$$\Pr(\boldsymbol{S_j}|H_1) = \begin{cases} \Pr(\boldsymbol{S_U}|\boldsymbol{S_K}, \boldsymbol{S_P}, H_1) & \{\boldsymbol{S_P} \subseteq \boldsymbol{S_j}\} \text{ and } \{\boldsymbol{S_K} \subseteq \boldsymbol{S_j}\} \\ 0 & \text{otherwise} \end{cases} \tag{4a}$$

Similarly

$$\Pr(\boldsymbol{S_{j'}}|H_2) = \begin{cases} \Pr(\boldsymbol{S_U}|\boldsymbol{S_K}, H_1) & \boldsymbol{S_K} \subseteq \boldsymbol{S_{j'}} \\ 0 & \text{otherwise} \end{cases} \tag{4b}$$

i.e. if the genotypes of the known contributors are present in the genotype set then the probability of the whole set is equal to the genotype probabilities of the unaccounted for genotypes (based on allele probabilities).

Note that $\boldsymbol{S_j}$ is an ordered set. This means that $\boldsymbol{S_P}$, $\boldsymbol{S_K}$ and $\boldsymbol{S_U}$ align with positions in the set. For an N person mixture there are $N!$ alignments (contributor orders) of $\boldsymbol{S_P}$, $\boldsymbol{S_K}$ and $\boldsymbol{S_U}$ to consider that may fulfil the condition $\{\boldsymbol{S_P} \subseteq \boldsymbol{S_j}\}$ and $\{\boldsymbol{S_K} \subseteq \boldsymbol{S_j}\}$.

Example 1. Simple two person mixture

It may be worthwhile considering an example at this point. Consider what may be the simplest mixed DNA profile, one that can be explained as coming from two persons reasonably assumed to be the two offenders. We initially consider only one locus.

The single accused (POI) has genotype G_p=[13,14] (note we switch from S to G terminology and unbold here to indicate that a single contributor's genotype is being considered). We consider the propositions:

H_1: the DNA is a mixture of the POI and an unknown.
H_2: the DNA is a mixture of two unknowns.

As is well understood, the genotypes that can make up the peaks seen in Fig. 1 are

$\boldsymbol{S_1}$ [13,14]:[16,17]	$\boldsymbol{S_4}$ [14,16]:[13,17]
$\boldsymbol{S_2}$ [13,16]:[14,17]	$\boldsymbol{S_5}$ [14,17]:[13,16]
$\boldsymbol{S_3}$ [13,17]:[14,16]	$\boldsymbol{S_6}$ [16,17]:[13,14]

We assume that the parameters suggest that the contributor in the first position, P_1, in the ordered set has more template than the person in the second position, P_2. It is to be expected that the weight of set one is close to one ($w_1 \approx 1$) and the others approach zero, either due to heterozygote imbalances (for S_2 to S_5) or an incompatibility of mixture proportions (for S_6, recall we select one of the two redundant solutions). Remembering that weights are normalised, a single genotype with w=1 is simply stating that there is only one genotype set that reasonably explains the observed data. Although we have shown only one locus, real applications will be at many loci. P_1 and P_2 retain their order across loci. We indicate the single person genotype at contributor position P_n in the genotype set $\boldsymbol{S_j}$ with a left superscript, so ${}^1S_1$=[13,14], ${}^2S_1$=[16,17], ${}^1S_2$=[13,16] etc.

Note that there are no known contributors in this example (i.e. contributors that are assumed to b e present in both H_1 and H_2) so $S_K = \varnothing$. Looking at the genotype sets described above, under H_1 the POI can occupy either P_1 or P_2, as the genotype G_p=[13,14] is present in both P_1 (${}^1S_1$) and P_2 (${}^2S_6$).

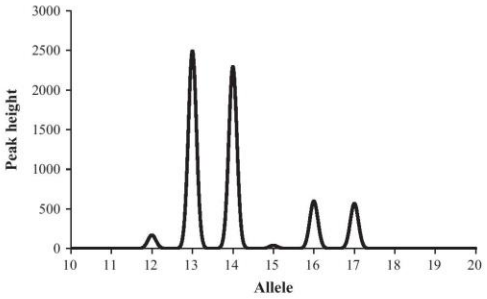


Fig. 1. Example of an electropherogram showing a single locus of a two person mixture.

This suggests that

$$\sum_j w_j \Pr(\boldsymbol{S_j}|G_P,H_1) = 1 \times \Pr(\boldsymbol{S_1}|G_P,H_1) + 0 \times \Pr(\boldsymbol{S_6}|G_P,H_1) = \Pr(\boldsymbol{S_1}|G_P,H_1)$$

End of Example 1.

Eqs. (2)–(4) apply at the whole-profile level, i.e. weights are full profile weights and genotype sets are the sets across all loci. It is uncommon to have such whole-profile information and much more common to consider evidence at the locus level. Including locus information in Eq. (2)

$$LR_C = \frac{\prod_l \sum_j w_j^l \Pr(\boldsymbol{S_j^l}|H_1)}{\prod_l \sum_{j'} w_{j'}^l \Pr(\boldsymbol{S_{j'}^l}|H_2)}$$

where l specifies a locus.

To consider contributor orders using data at an individual locus level a dependency of genotype sets on contributor order must be introduced. The LR given in Eq. (2) then becomes

$$LR_{PP1} = \frac{\sum_{i=1}^{N!} \prod_l \sum_j w_j^l \Pr(\boldsymbol{S_j^l}|\mathbb{C}_i,H_1) Pr(\mathbb{C}_i)}{\sum_{i=1}^{N!} \prod_l \sum_{j'} w_{j'}^l \Pr(\boldsymbol{S_{j'}^l}|\mathbb{C}_i,H_2) Pr(\mathbb{C}_i)} \tag{5}$$

There is no need to specify a contributor order for weights as they apply to multi-contributor genotype sets regardless of which order the known contributors are being considered (i.e. in Examples 1, $w1 \approx 1$ regardless of whether $\mathbb{C}_1$ or $\mathbb{C}_2$ is being considered). If we consider equal priors for each contributor order then the $Pr(\mathbb{C}_i)$ can be dropped from Eq. (5) to give

$$LR_{PP1} = \frac{\sum_{i=1}^{N!} \prod_l \sum_j w_j^l \Pr(\boldsymbol{S_j^l}|\mathbb{C}_i,H_1)}{\sum_{i=1}^{N!} \prod_l \sum_{j'} w_{j'}^l \Pr(\boldsymbol{S_{j'}^l}|\mathbb{C}_i,H_2)} \tag{6}$$

To ensure that contributor order is maintained between loci, Eq. (4a) becomes

$$\Pr(\boldsymbol{S_j^l}|\mathbb{C}_i,H_1) = \begin{cases} \Pr(\boldsymbol{S_u^l}|\boldsymbol{S_K^l},\boldsymbol{S_p^l},\mathbb{C}_i,H_1) & \left\{\boldsymbol{S_p^l} = {}^n\boldsymbol{S_j^l}\right\} \text{ and } \left\{\boldsymbol{S_K^l} = {}^{n'}\boldsymbol{S_j^l}\right\} \\ 0 & \text{otherwise} \end{cases} \tag{7a}$$

To be included as a non-zero term the known contributors and POI must not only be present in $\boldsymbol{S_j^l}$ (as specified in Eq. (4a)) but also be present in the positions dictated by the specified contributor order across all loci. If there are multiple POIs or known contributors, all of their genotypes must be present in the genotype sets at the positions dictated by the contributor order.

Considering locus and contributor order Eq. (4b) becomes

$$\Pr(\boldsymbol{S_j^l}|\mathbb{C}_i,H_2) = \begin{cases} \Pr(\boldsymbol{S_u^l}|\boldsymbol{S_K^l},\mathbb{C}_i,H_2) & \boldsymbol{S_K^l} = {}^{n'}\boldsymbol{S_j^l} \\ 0 & \text{otherwise} \end{cases} \tag{7b}$$

Example 2. The difference between LR_{PP1} and LR_{PP2}

We consider the scenario stated in Example 1 more generally now under three situations; where the profile is fully resolvable, partially resolvable, and unresolvable. For each situation the same six possible genotype sets are considered but this time given weights to reflect the scenario (see Table 1).

Under H_2 we have two unknown contributors, G_{U1} and G_{U2}. The probability of the observed mixture given the ordered set $G_{U1}=[13,14]$; $G_{U2}=[16,17]$ is equivalent to the probability of the observed mixture given the ordered set $G_{U2}=[13,14]$; $G_{U1}=[16,17]$. This leads us to the concept of degeneracy in the ordering of the unknowns. If N unknown individuals are being considered under H_2 as the source of DNA of an N person mixture then $N!$ equal probabilities will be summed in the denominator of LR_{PP1}. This degeneracy simplifies Eq. (6) to

$$LR_{PP1} = \frac{\sum_{i=1}^{N!} \prod_l \sum_j w_j^l \Pr(\boldsymbol{S_j^l}|\mathbb{C}_i,H_1)}{N! \prod_l \sum_{j'} w_{j'}^l \Pr(\boldsymbol{S_{j'}^l}|H_2)}$$

As we are considering only a single locus the locus term can also be dropped from the equation. The two contributor orders in our example of two person mixture (Fig. 1) give

$$LR_{PP1} = \frac{w_1 \Pr(\boldsymbol{S_1}|G_P,H_1) + w_6 \Pr(\boldsymbol{S_6}|G_P,H_1)}{2 \sum_{j'=1}^{6} w_{j'} \Pr(\boldsymbol{S_{j'}}|G_P,H_2)}$$

Example 1 showed $\Pr(\boldsymbol{S_1}|G_P,H_1) = \Pr(\boldsymbol{S_6}|G_P,H_1)$, which we call $\Pr(\boldsymbol{S_1}|H_1)$ giving

$$LR_{PP1} = \frac{(w_1+w_6)\Pr(\boldsymbol{S_1}|G_P,H_1)}{2 \sum_{j'=1}^{6} w_{j'} \Pr(\boldsymbol{S_{j'}}|G_P,H_2)}$$

which can be related to LR_{PP2} by

$$LR_{PP1} = \frac{w_1 \Pr(\boldsymbol{S_1}|G_P,H_1)}{R \sum_{j'=1}^{6} w_{j'} \Pr(\boldsymbol{S_{j'}}|G_P,H_2)} = \frac{LR_{PP2}}{R}$$

where

$$R = \frac{2}{(1+(w_6/w_1))} \tag{8}$$

Fully resolvable:

$w_1=1$ and $w_6=0$ giving $R=2$ and $LR_{PP1}=LR_{PP2}/2$, which if enumerated would give the LR in Eq. (1).

Unresolved:

$w_1=w_6$ giving $R=1$ and $LR_{PP1}=LR_{PP2}$.

Partially resolved:

$w_1=0.59$ and $w_6=0.01$ giving $R=1.97$ and $LR_{PP1}=LR_{PP2}/1.97$, which shows the difference that the choice of propositions will have on the resulting LR when there is partial resolution of the two contributor components.

Table 1
Weights for genotype sets under three scenarios (note the weights sum to one for each scenario by design).

Genotype set (j)	Genotype set (S_j)		Weights (w_j)		
	^1S_j	^2S_j	Fully resolved	Partially resolved	Unresolved
1	[13,14]	[16,17]	1	0.59	0.167
2	[13,16]	[14,17]	0	0.10	0.167
3	[13,17]	[14,16]	0	0.10	0.167
4	[14,16]	[13,17]	0	0.10	0.167
5	[14,17]	[13,16]	0	0.10	0.167
6	[16,17]	[13,14]	0	0.01	0.167

For the example given in Fig. 1 the level of resolution, and its effect it will have on R (Eq. (8)) can be seen in Fig. 2. **End of Example 2.**

Example 3. Multiple loci

Imagine two loci, both with four peaks as in Example 1, (Fig. 3). We start again from

$$LR_{PP1}=\frac{\sum_{i=1}^{N!}\prod_{l}\sum_{j}w_j^l\Pr(\boldsymbol{S_j^l}|\mathbb{C}_i,H_1)}{N!\prod_{l}\sum_{j'}w_{j'}^l\Pr(\boldsymbol{S_{j'}^l}|H_2)}$$

In Example 2 we showed that when $G_P^1=[13,14]$ and considering only the first locus:

$$LR_{PP1}=\frac{\Pr(\boldsymbol{S_1}|G_P,H_1)}{2\sum_{j'=1}^{6}w_{j'}\Pr(\boldsymbol{S_{j'}}|G_P,H_1)}$$

Now consider two scenarios for locus 2 as seen in Fig. 3:

1) $G_P^2=[13,14]$. In this example the POI contains the major alleles at loci 1 and 2, and is therefore expected to yield a $LR>1$ when considering both loci.
2) $G_P^2=[16,17]$. In this example then the POI contains the major alleles at locus 1 and the minor alleles at allele 2, and so would be expected to have $LR=0$ when considering both loci.

Scenario 1

We showed in Example 2 that $\sum_{i=1}^{N!}\prod_l\sum_j w_j^l\Pr(\boldsymbol{S_j^l}|\mathbb{C}_i,H_1)=(w_1^1+w_6^1)\Pr(\boldsymbol{S_1^l}|G_p^1,H_1)$, but this was considering only one locus. We now consider two loci with information provided at a locus by locus level, and so contributor order must be maintained across loci. In $\mathbb{C}_1$ only one genotype set fulfils $S_p^l={}^nS_j^l$ from Eq. (7a), and that is genotype set one, as $S_p^1={}^1S_1^1$. Similarly for locus 2 the only genotype set that fulfils $S_p^l={}^nS_j^l$ is again genotype set one, $S_p^2={}^1S_1^2$. In $\mathbb{C}_2$ the genotype sets that fulfil $S_p^l={}^nS_j^l$ at loci 1 and 2 are S_6^1 and S_6^2 respectively. Therefore $\sum_{i=1}^{N!}\prod_l\sum_j w_j^l\Pr(\boldsymbol{S_j^l}|\mathbb{C}_i,H_1)$ can be calculated by

$$w_1^1\Pr(\boldsymbol{S_1^l}|G_p^1,H_1)w_1^2\Pr(\boldsymbol{S_1^2}|G_p^2,H_1)+w_6^1\Pr(\boldsymbol{S_6^l}|G_p^1,H_1)w_6^2\Pr(\boldsymbol{S_6^2}|G_p^2,H_1)$$

Simplifying as in Example 2 yields

$$=(w_1^1w_1^2+w_6^1w_6^2)Pr(\boldsymbol{S_1^1}|G_p^1,H_1)Pr(\boldsymbol{S_1^2}|G_p^2,H_1)$$

And so

$$LR_{PP1}=\frac{(w_1^1w_1^2+w_6^1w_6^2)Pr(\boldsymbol{S_1^1}|G_p^1,H_1)Pr(\boldsymbol{S_1^2}|G_p^2,H_1)}{N!\prod_l\sum_{j'=1}^{6}w_{j'}^l\Pr(\boldsymbol{S_{j'}^l}|G_p^l,H_2)}$$

$$=\frac{w_1^1w_1^2Pr(\boldsymbol{S_1^1}|G_p^1,H_1)Pr(\boldsymbol{S_1^2}|G_p^2,H_1)}{R\prod_l\sum_{j'=1}^{6}w_{j'}^l\Pr(\boldsymbol{S_{j'}^l}|G_p^l,H_2)}=\frac{LR_{PP2}}{R}$$

where when considering two loci:

$$R=\frac{2}{\left(1+\frac{w_6^1w_6^2}{w_1^1w_1^2}\right)}$$

We make the simplifying assumption for this example that the product of the weights for a genotype set across all loci ($\prod_{l=1}^{L}w_j^l$) is approximately equal to the average weight ($\overline{w_j}$) raised to the power of the number of loci, $\prod_{l=1}^{L}w_j^l\approx(\overline{w_j})^L$ giving

$$R=\frac{2}{1+\left(\frac{w_6}{w_1}\right)^L}$$

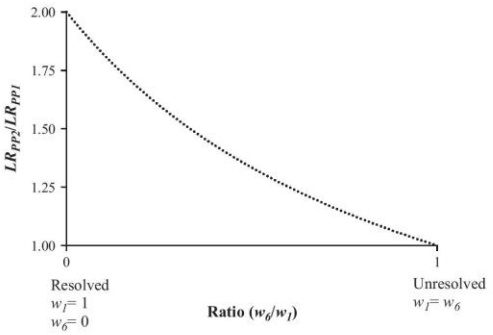


Fig. 2. Difference that the choice of propositions will have on the reported LR using the example seen in Fig. 1, considering level of resolution in P_1 and P_2.

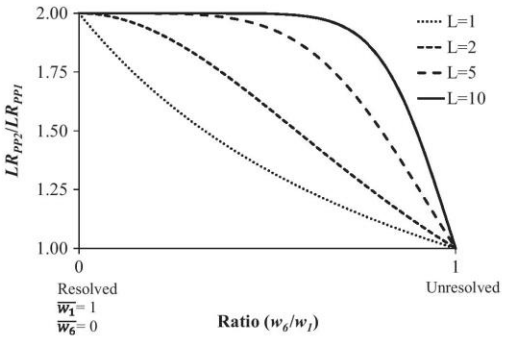


Fig. 4. The effect that different propositions will have on the reported LR using the example seen in Fig. 1, considering the level of resolution in P_1 and P_2 and multiple loci (L).

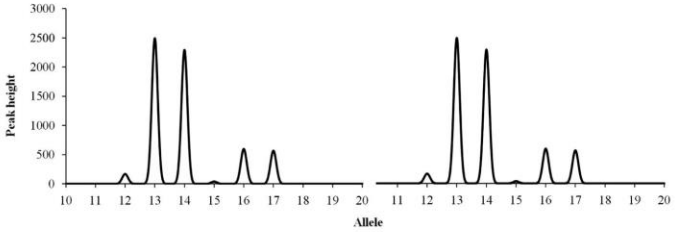


Fig. 3. Example of an electropherogram showing two loci of a two person mixture.

Fig. 2 can be reproduced, this time considering multiple loci (Fig. 4). The range for R is always from 1 to 2 for this two person mixture example, but the rate at which it decreases varies with number of loci.

Scenario 2

Carrying out the same process, $\sum_{l=1}^{N!}\prod_l\sum_j w_j^l \Pr(\boldsymbol{S_j^l}|\mathbb{C}_i,H_1)$ can be calculated by

$$w_1^1\Pr(\boldsymbol{S_1^1}|G_p^1,H_1)w_6^2\Pr(\boldsymbol{S_6^2}|G_p^2,H_1)+w_6^1\Pr(\boldsymbol{S_6^1}|G_p^1,H_1)w_1^2\Pr(\boldsymbol{S_1^2}|G_p^2,H_1)$$
$$=(w_1^1w_6^2+w_6^1w_1^2)Pr(\boldsymbol{S_1^1}|G_p^1,H_1)Pr(\boldsymbol{S_6^2}|G_p^2,H_1)$$

For a fully resolved profile we know $w_6\approx 0$ and so $(w_1^1w_6^2+w_6^1w_1^2)=0$ and $LR_{PP1}=0$ as expected.

End of Example 3.

Assuming contributors to a DNA profile (i.e. nominating individuals in $\boldsymbol{S_K}$) has the effect of reducing the number of effective contributor orders. For example if a four person mixture was analysed with three contributors being assumed under H_1 and H_2, this would leave a single unaccounted contributor position to which POIs could be compared. Thinking of this in the $N!$ space, there would be 23 contributor orders that would result in a probability of 0 under H_1 and H_2 and one that would result in a non-zero value (assuming the POI was not excluded). The 23 zero probabilities would mean LR_{PP1} has the same value that it would if it were considered as having only a single contributor order. In the mathematical form the above example is saying

$$LR_{PP1}=\frac{\prod_l\sum_j w_j^l\Pr(\boldsymbol{S_j^l}|\mathbb{C}_1,H_1)+\sum_{i=2}^{24}\prod_l\sum_j w_j^l\Pr(\boldsymbol{S_j^l}|\mathbb{C}_i,H_1)}{\prod_l\sum_{j'} w_{j'}^l\Pr(\boldsymbol{S_{j'}^l}|\mathbb{C}_1,H_2)+\sum_{i=2}^{24}\prod_l\sum_{j'} w_{j'}^l\Pr(\boldsymbol{S_{j'}^l}|\mathbb{C}_i,H_2)}$$
$$=\frac{\prod_l\sum_j w_j^l\Pr(\boldsymbol{S_j^l}|\mathbb{C}_1,H_1)+0}{\prod_l\sum_{j'} w_{j'}^l\Pr(\boldsymbol{S_{j'}^l}|\mathbb{C}_1,H_2)+0}$$
$$=\frac{\prod_l\sum_j w_j^l\Pr(\boldsymbol{S_j^l}|\mathbb{C}_1,H_1)}{\prod_l\sum_{j'} w_{j'}^l\Pr(\boldsymbol{S_{j'}^l}|\mathbb{C}_1,H_2)}$$

3. Conclusion

We have reworked the familiar 'factor of two' phenomenon in a context amenable to use in continuous models. Unlike the classic 'factor of two' phenomenon, which deals with discreet and separate evidence electropherograms, the phenomenon in the context of components of a single mixed profile can be partially or completely unresolved. This leads to the requirement of a different treatment for the issue. The examples given are simple, but very specific. In general the effect that the different propositions will have on the calculated *LR* depends on:

- the number of contributors to the evidence;
- the number and genotypes of the POI(s) and known contributors;
- the level of resolution in the contributor components in the evidence profile;
- the specific weights of the genotype sets; and
- the number of loci being considered.

We discuss propositions at the sub–sub-source level and the sub-source level. All comparisons must eventually be placed at the offence level. This means that the propositions considered here are subordinate to the source, activity and finally offence levels. This last set of propositions, the offence level ones, are usually left for the court but it is important to consider that there are many considerations between the sub-source propositions and the full context of the case. Bayes networks have been demonstrated to be of significant use in developing these considerations (Gittelson et al., 2012, 2013).

An analyst that reports an *LR* considering the evidence at the sub-source level compares genotypes and unknowns to all contributor positions within a profile and so considers the propositions:

H_1: The POI is one of the individuals in the mixture.
H_2: Unknown people make up the mixture.

When contributor order is not considered (typically the maximum *LR* value that is produced out of the $N!$ contributor alignments) the *LR* reports at the sub–sub-source level as it considers the propositions:

H_1: The POI is the donor of a specific portion of DNA in the mixture.
H_2: An unknown person is the donor of the specific portion of DNA in the mixture.

Eq. (6) has been derived and can be used to calculate an *LR*, using locus by locus data, at the sub-source level, taking into account the level of resolution between contributors within the mixture.

In many instances the difference between LR_{PP1} and LR_{PP2} would have no practical impact at an offence level. Despite this fact we believe that if given the ability to calculate an *LR* using the most statistically correct theory, analysts are obliged to do so. At the very least, an analyst should have the knowledge of the magnitude of the effect on the *LR* that their assumptions have, so that an appropriate adjustment can be made. We suggest that the appropriate set of propositions for use in *LR* calculations is PP_1, but recognise that this can be technically difficult. We suggest that scientists have the following options (listed in order of preference):

(a) Calculate and report *LR*s at the sub-source level using PP_1.
(b) Use a simplified approach and simply divide the *LR* for an N person profile by $N!$.
(c) Calculate and report *LR*s at the sub–sub-source level, but outline in reports what the propositions are that they are actually using and the limitations of this approach.

This is very close to the option list given by Champod and Buckleton in Buckleton et al. (2004, pp. 255). Note that option (a) does not require that the same number of contributors be chosen in H_1 and H_2 whereas options (b) and (c) do. In many instances this will not pose a problem beyond what many forensic laboratories currently face with limitations in the software being used. For laboratories that wish to carry out calculations that consider a different number of contributors in each proposition then a modification can be made to preference (b) which increases complexity, but also versatility:

If U_1 is the number of unknown individuals in H_1 and U_2 is the number of unknowns in H_2 then the *LR* at the sub–sub-source level can be multiplied by $U_1!/U_2!$ to obtain the *LR* at the sub-source level.

Note that this modified option (b) can still be used when the number of contributors between the two propositions is the same.

As propositions are considered at higher levels in the hierarchy of propositions there are additional factors that must be considered, which may dominate the *LR*. Source level propositions link an individual to a biological source and so consideration of laboratory

contamination and error rates become important. If contamination or error rates are high then this will dominate the *LR*. Even at the sub-source level, consideration of typing errors can be taken into account when calculating the *LR*; we direct the reader to Balding (2005) for a mathematical treatment of typing errors.

Acknowledgements

This work was supported in part by Grant 2011-DN-BX-K541 from the US National Institute of Justice. Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the U.S. Department of Justice. We would like to thank Johanna Veth and two anonymous reviewers for their helpful comments that improved this paper. We would also like to thank Ian Evett for helpful contributions to this work.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.jtbi.2014.08.021.

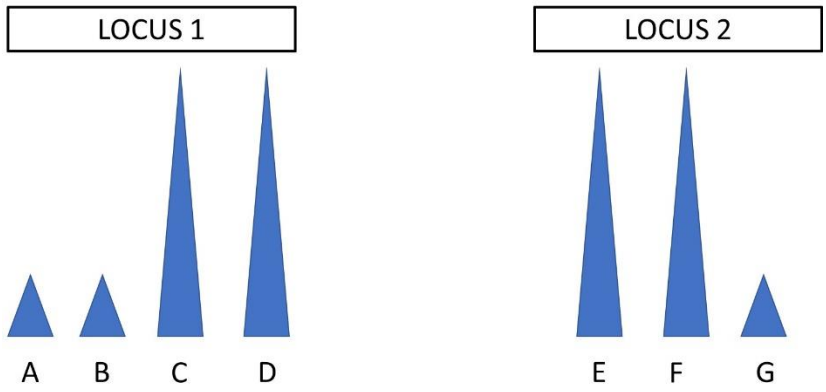
References

Balding, D.J., 2005. Weight-of-evidence for Forensic DNA ProfilesJohn Wiley and Sons, Chichester.
Balding, D.J., 2013. Evaluation of mixed-source, low-template DNA profiles in forensic science. Proc. Natl. Acad. Sci. USA 110, 12241–12246.
Balding, D.J., Buckleton, J., 2009. Interpreting low template DNA profiles. Forensic Sci. Int. Genet. 4 (1), 1–10.
Berger, C.E.H., et al., 2011. Evidence evaluation: a response to the court of appeal judgment in R v T. Sci. Justice 51 (2), 43–49.
Buckleton, J.S., Triggs, C.M., Walsh, S.J., 2004. DNA Evidence. CRC Press, Boca Raton, Florida.
Cook, R., et al., 1998a. A hierarchy of propositions: Deciding which level to address in casework. Sci. Justice 38 (4), 231–240.
Cook, R., et al., 1998b. A model for case assessment and interpretation. Sci. Justice 38 (3), 151–156.
Cowell, R.G., Lauritzen, S.L., Mortera, J., 2007. Identification and separation of DNA mixtures using peak area information. Forensic Sci. Int. 166 (1), 28–34.
Evett, I.W., 1987. On meaningful questions: a two-trace transfer problem. J. Forensic Sci. Soc. 27, 375–381.
Evett, I.W., Gill, P.D., Lambert, J.A., 1998. Taking account of peak areas when interpreting mixed DNA profiles. J. Forensic Sci. 43 (1), 62–69.
Evett, I.W., Jackson, G., Lambert, J.A., 2000. More on the hierarchy of propositions: exploring the distinction between explanations and propositions. Sci. Justice 40 (1), 3–10.
Evett, I.W., et al., 2002. Interpreting small quantities of DNA: the hierarchy of propositions and the use of Bayesian networks. J. Forensic Sci. 47 (3), 520–530.
Gill, P., et al., 1998. Interpreting simple STR mixtures using allelic peak areas. Forensic Sci. Int. 91, 41–53.
Gittelson, S, et al., 2012. Bayesian networks and the value of the evidence for the forensic two-trace transfer problem. J. Forensic Sci. 57 (5), 1199–1216.
Gittelson, S, et al., 2013. Modeling the forensic two-trace problem with Bayesian networks. Artif. Intell. Law 21 (2), 221–252.
Haned, H., 2011. Forensim: an open-source initiative for the evaluation of statistical methods in forensic genetics. Forensic Sci. Int.: Genet. 5 (4), 265–268.
Jackson, G., et al., 2006. The nature of forensic science opinion – a possible framework to guide thinking and practice in investigations and in court proceedings. Sci. Justice— J. Forensic Sci. Soc. 46 (1), 33–44.
Lohmueller, K., Rudin, N., 2013. Calculating the weight of evidence in low-template forensic DNA casework. J. Forensic Sci. 58 (1), 234–259.
Meester, R., Sjerps, M., 2003. The evidential value in the DNA database search controversy and the two-stain problem. Biometrics 59 (3), 727–732.
Perlin, M.W., Sinelnikov, A., 2009. An information gap in DNA evidence interpretation. PLoS One 4 (12), e8327.
Perlin, M.W., et al., 2011. Validating TrueAllele® DNA mixture interpretation. J. Forensic Sci. 56, 1430–1447.
Taylor, D., Bright, J.-A., Buckleton, J., 2013. The interpretation of single source and mixed DNA profiles. Forensic Sci. Int.: Genet. 7 (5), 516–528.
Triggs, C.M., Buckleton, J., 2003. The two trace transfer problem revisited. Sci. Justice 43 (3), 127–134.
Triggs, C.M., Buckleton, J.S., 2004. Comment on "Why the effect of prior odds should accompany the likelihood ratio when reporting DNA evidence". In: Meester, R., Sjerps, M. Law, Probability and Risk, vol. 3, 1, pp. 73–82.

3.2 – Clarification

Point 1: clarification on the contributor order nomenclature using a multi-locus example

Consider a system that yields a posterior distribution for each parameter that is used to describe a DNA profile. When presented with a DNA profile that has DNA contributed in unequal amounts, then the posterior distributions relating to amount of DNA will be different for each contributor. Consider the following example of a DNA profile at two loci:



We can see that at each locus there are high peaks and low peaks, and we might suggest that there is a major contributor and a minor contributor of DNA to this sample. We would expect that a contributor dependant template parameter would have a distribution with a higher mean for the major contributor then the minor contributor. Consider that in this example the peak heights between the two contributors are divergent enough that the probability of a high peak partnering with a low peak is effectively 0. Also, consider that the heights of all peaks are such that we are satisfied we are seeing all contributor peaks (i.e. there has been no dropout of any of the contributor's alleles). Under these conditions, at locus 1, we would have acceptable genotypes:

[A,B] and [C,D]

And for locus 2 we have

[E,F], [E,G], [F,G] and [G,G]

Note that some of these genotypes would correspond to the major contributor and others to the minor contributor, but I have not specified any order in the sets given. Also note that template DNA amount acts across loci, i.e. the major contributor at one locus, must also be the major contributor at another locus (at this stage ignoring situations where degradation has acted to different degrees on the two contributors so that the major at low molecular weight loci could become the minor at high molecular weight loci). Considering multi-locus genotypes, we would therefore not expect a donor of DNA to the sample to have:

[A,B] & [E,F]

But could have:

[A,B] & [E,G]

It is useful to separate the genotypes that each contributor could have. The table below provides this information for the profile above:

	Contributor 1 genotypes	Contributor 2 genotypes
Locus 1	[C,D]	[A,B]
Locus 2	[E,F]	[E,G], [F,G] and [G,G]

In the table above the ordering is that contributor 1 is the major contributor and contributor 2 is the minor contributor, although this is not important (i.e. I could have just as readily switched the genotypes for the two contributors as long as I maintained the same sets of genotypes at both loci for each contributor).

This introduces the concept of an order in genotype sets. Consider *J* sets of *N* genotypes at locus *l*, $\mathbf{S}_{\mathbf{j}}^{\mathbf{l}} = \left\{ {}^{1}G_j^l, ..., {}^{N}G_j^l \right\}$. In the example, there is one genotype set at locus 1:

$$\mathbf{S}_1^1 = \left\{ [C, D], [A, B] \right\}$$

and three at locus 2:

$$\mathbf{S}_1^2 = \left\{ [E, F], [E, G] \right\}$$

$$\mathbf{S}_2^2 = \left\{ [E, F], [F, G] \right\}$$

$$\mathbf{S}_3^2 = \left\{ [E, F], [G, G] \right\}$$

The order of genotypes in the set is given by convention that position 1 is the major and position 2 is the minor. These genotype sets have associated with them posterior distributions for parameters.

Consider now that we wish to compare a POI who has genotype [A,B] at locus 1 and [F,G] at locus 2. We could compare them to the mixed DNA profile using propositions:

Hp: The POI is contributor 1 and an unknown is contributor 2

Hd: Both contributors are unknowns

As [A,B] does not appear in the list of genotypes for set position 1, and [F,G] does not appear in the genotypes for set position 2 and so the *LR* calculated using the above propositions would be 0. However, an LR calculated using the propositions:

Hp: The POI is contributor 2 and an unknown is contributor 1

Hd: Both contributors are unknowns

The we would expect some support for the inclusion of the POI as a donor to the mixture.

So, we find that not only is the order of genotypes in genotype sets important, but also the contributor positions to which references are compared. We term this latter the contributor order. We signify contributor order by $\mathbb{C}_{N,c}$, where *N* is the total number of contributors to the evidence DNA profile and *c* is the contributor order. For a single contributor profile there is a

single position with which the POI can be compared and so only one contributor order exists, $\mathbb{C}_{1,1}$, where $\left\{\mathbb{C}_{1,1}:(\mathrm{POI}\equiv {}^{1}G)\right\}$ and I use $\equiv$ to signify a comparison between a person from the proposition (typically POI or an unknown) and a genotype set position. For a two-person profile there are two contributor orders $\mathbb{C}_{2,1}$ and $\mathbb{C}_{2,2}$. In the first 2-person contributor order ($\mathbb{C}_{2,1}$) the POI is compared with the genotypes in position one and the unknown (U_1 for 'unknown 1') is compared to the genotypes in position two $\left\{\mathbb{C}_{2,1}:(\mathrm{POI}\equiv {}^{1}G,\mathrm{U}_1\equiv {}^{2}G)\right\}$. The order is reversed in contributor order two $\left\{\mathbb{C}_{2,2}:(\mathrm{U}_1\equiv {}^{1}G,\mathrm{POI}\equiv {}^{2}G)\right\}$. Note that the genotype set position always corresponds to the position that the elements of the contributor order are presented. From this point forward, I drop the explicit statement of genotype set position. For three person profile there are six orders $\left\{\mathbb{C}_{3,1}:\left(\mathrm{POI},\mathrm{U}_1,\mathrm{U}_2\right)\right\}$, $\left\{\mathbb{C}_{3,2}:\left(\mathrm{POI},\mathrm{U}_2,\mathrm{U}_1\right)\right\}$, $\left\{\mathbb{C}_{3,3}:\left(\mathrm{U}_1,\mathrm{POI},\mathrm{U}_2\right)\right\}$, $\left\{\mathbb{C}_{3,4}:\left(\mathrm{U}_1,\mathrm{U}_2,\mathrm{POI}\right)\right\}$, $\left\{\mathbb{C}_{3,5}:\left(\mathrm{U}_2,\mathrm{POI},\mathrm{U}_1\right)\right\}$ and $\left\{\mathbb{C}_{3,6}:\left(\mathrm{U}_2,\mathrm{U}_1,\mathrm{POI}\right)\right\}$. For an *N* person profile there are *N!* contributor orders.

3.3: Treating parameters in the LR as distributions using highest posterior density

When STRmix™ was introduced into forensic laboratories, the forensic community grappled with the introduction of a new source of variability within the generation of the LR, namely the variability due to using a stochastic process such as MCMC. The legal community also had difficulty with the idea that each analysis (even of the same DNA profile) would produce a different result. For too long had 'reproducibility' been synonymous with 'reliability'. In fact, the reproducibility of DNA results had never really existed. Every step of generating a DNA profile (the amount of DNA sampled from an item, the amount of DNA recovered from the sample, the functioning of the PCR and the functioning of the electrophoresis) is subject to stochastic variation. Even the previous methods of LR calculation only gave the same answer due to the assumptions and simplifications of the models. There was never any guarantee that the LRs from these early systems were producing accurate values, and users were mistaking the precision with accuracy. For those readers that are familiar with stochastic systems and random number generation, imagine absolute reproducibility could be forced within an MCMC system by supplying the same random number seed. All would agree that doing this should not be viewed as an improvement. Setting the seed will not have improved accuracy or reliability and only achieved reproducibility artificially. Introducing a continuous DNA interpretation system did not introduce more uncertainty into the LR assignment, merely removed the illusion of stability that the simple systems portrayed.

A natural question arising from the use of MCMC was "*how much could the LR vary from run MCMC to run?*" and "*what factors will cause it to vary?*". For some time, the forensic community had been accustomed to the idea of accounting for sampling variation in allele frequencies by providing a confidence (or credible) interval on the LR. These allele frequencies are used to calculate the rarity of a DNA profile, and are based on survey of a population, which are finite and so subject to random variation depending on who happened to be included in the survey. The method for producing such a credible interval is described in the publication in section 2.6 using the highest posterior density (HPD). The variability in the LR produced by STRmix™ lead to a body of work that included additional factors (additional to allele sampling variation) in the HPD credible interval. Specifically, a method was devised to take into account the amount of variability expected from the stochasticity of MCMC by using a resampling method based on the effective sample size of the analysis. This work is described in the publication in section 3.3.

In the work on LR variability, the genetic model was also extended to include the possibility that the true offender (if not the suspect) may be a relative of the suspect. This had been considered in the past, but not for the complex, mixture based, LR calculations that were being performed in STRmix™. The inclusion of relatives as alternate DNA donors produced a 'unified' LR, which is also reported in the publication in section 3.3 and takes steps towards the full Bayesian approach spoken about in chapter 1. To achieve this latter task required that mathematics be developed that allowed the consideration of relatives of a suspect contributing DNA to a complex mixture. This work is described in the publication in section 3.4.

Manuscript: An illustration of the effect of various sources of uncertainty on DNA likelihood ratio calculations. D Taylor, JA Bright, J Buckleton, J Curran. (2014) Forensic Science International: Genetics 11, 56-63 – *Cited 10 times*

Statement of novelty: This work explores various parameters used in the LR calculation and extends existing theory to show how they can be considered as distributions rather than point values (which is how they had been treated up until this point).

My contribution: I was main author, main theorist and carried out all simulation work for this paper.

Research Design / Data Collection / Writing and Editing = 85% / 100% / 70%

Additional comments:

Forensic Science International: Genetics 11 (2014) 56–63

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

An illustration of the effect of various sources of uncertainty on DNA likelihood ratio calculations

D. Taylor [a,*], J-A. Bright [b], J. Buckleton [b], J. Curran [c]

[a] Forensic Science South Australia, 21 Divett Place, SA 5000, Australia
[b] ESR Ltd, Private Bag 92021, Auckland 1142, New Zealand
[c] Department of Statistics, University of Auckland, Private Bag 92019, Auckland 1142, New Zealand

ARTICLE INFO

Article history:
Received 29 August 2013
Received in revised form 28 January 2014
Accepted 3 February 2014

Keywords:
DNA interpretation
Sampling uncertainty
MCMC
HPD
Continuous methods
Relatives

ABSTRACT

A typical assessment of the strength of forensic DNA evidence is based on a population genetic model and estimated allele frequencies determined from a population database. Some experts provide a confidence or credible interval which takes into account the sampling variation inherent in deriving these estimates from only a sample of a total population. This interval is given in conjunction with the statistic of interest, be it a likelihood ratio (LR), match probability, or cumulative probability of inclusion. Bayesian methods of addressing database sampling variation produce a distribution for the statistic from which the bound(s) of the desired interval can be determined.

Population database sampling uncertainty represents only one of the sources of uncertainty that affects estimation of the strength of DNA evidence. There are other uncertainties which can potentially have a much larger effect on the statistic such as, those inherent in the value of F_{st}, the weights given to genotype combinations in a continuous interpretation model, and the composition of the relevant population. In this paper we model the effect of each of these sources of uncertainty on a likelihood ratio (LR) calculation and demonstrate how changes in the distribution of these parameters affect the reported value. In addition, we illustrate the impact the different approaches of accounting for sampling uncertainties has on the LR for a four person mixture.

1. Introduction

The likelihood ratio is the recommended statistic for the calculation of the weight of forensic DNA evidence [1]. It is the ratio of the probability of the observed evidence DNA profile (O) given each of two competing hypotheses, H_1 and H_2.

Traditional methods of DNA profile interpretation are described as binary. In a binary interpretation system weights of zero or one are used to either exclude or include genotype sets respectively. The weights represent a relative assignment of the probability density of the observed profile if it is from the proposed genotype combination. Hereafter we will refer to a probability density as a probability for simplicity and because the difference, although important, is not required here.

This assignment of relative probability is often guided by a set of heuristics that may include heterozygous balance, dropout, and mixture proportion [2]. A continuous interpretation model uses the quantitative information from an electropherogram such as peak heights, to calculate the probability of the peak heights given all possible genotype set combinations, S_j. A weight, w_j, can be defined as the normalised probability density of the observed evidence data (O) given the proposed genotype set combination, $\Pr(O|S_j)$.

Weight is a relatively new term for a concept that has been in use in DNA profile interpretation for some time. The variation in these weights assigned using a binary method of interpretation is difficult to quantify, as any variation will be from the differences arising in the interpretations between two, or more, analysts. Advances in research along with access to increased computer resource, have given practicing forensic scientists access to software which generate and apply continuous (as opposed to binary) weights. These weights are often estimated by Markov chain Monte Carlo (MCMC) methods [3–5] utilising peak height information and models of DNA profile behaviours. Continuous methods allow weights to be assigned any value between zero and one. In the frequentist paradigm, the weights are regarded as having a fixed, but unknown value. A reasonable frequentist procedure looks to use the data to provide an estimate of the weights along with the associated uncertainty in the estimates. In the Bayesian school of thought, the weights are regarded as random, with their behaviour described by a statistical distribution. Under either framework it is typical to only use the average weights when calculating the LR. However, in reality the weights

* Corresponding author. Tel.: +61 8 8226 7700; fax: +61 8 8226 7777.
E-mail address: Duncan.Taylor@sa.gov.au (D. Taylor).

1872-4973/$ – see front matter.
http://dx.doi.org/10.1016/j.fsigen.2014.02.003

are unknown, and it is useful to consider the effect of this uncertainty on the resulting LR calculations.

Some forensic practitioners calculate a credible interval (CI) or confidence interval that accounts for the sampling variation inherent in allele frequency estimation from a sample of a population of interest; namely a population database. Not all commentators believe that an assessment of sampling error is necessary. Brenner [6] makes explicit his doubts about the usefulness of assessing sampling uncertainty with the following challenge:

> "Will someone tell me, please, what rational difference it ever can make to know the confidence limits in addition to knowing the best point estimate? Specifically, can you give premises under which, for a fixed point estimate, the decision to convict or not to convict would depend on the size of the confidence interval?"

There is a lot of substance to Brenner's challenge. However these comments may not have taken full account of the cross examination process where any uncertainty or doubt is often explored at length. An analyst who has prepared for such a cross examination will likely be of more assistance to the court than one who chooses to answer "would it make any difference?" Furthermore, and perhaps more importantly, in our experience it is an accepted practice in adversarial systems that all reasonable uncertainty is conceded to the defendant.

Commenting on statistical evidence in general, rather than DNA in particular, Good stated [7]:

> "The court expects us to provide both an average based on our sample and some measure of the accuracy of our average."

Almost any measurement in science has an associated measure of uncertainty. Well prepared lawyers correctly investigate this avenue of questioning. In our experience, this typically has been approached by asking a question along the lines: "Is your database of 180 individuals big enough?"

Is there any reason why DNA evidence should be exempt from this line of questioning? The position advocating a consideration of sampling uncertainty is also taken by many authors [8–14]. In most cases, even with the inclusion of an estimate of sampling uncertainty, the final answer is not vastly different to the point or 'best' estimate.

One method for calculating a CI is the highest posterior density (HPD) [14–16]. The HPD method allows the calculation of interval bounds and uses a Dirichlet distribution to describe the variation in allele frequencies estimated from a population database. The HPD interval bounds are evaluated using a Monte Carlo method. In this method a large random sample is taken from the posterior distribution of the LR and the empirical sample quantiles are used as estimates of the bounds.

Whilst the HPD method has been applied to forensic LR calculations with respect to the variability inherent in allele frequency estimates, it can also readily account for the variance of other parameters impacting on the LR, namely F_{st} (commonly called theta, or the coancestry coefficient), genotype set weights and population composition. We will investigate the effect of each of these factors of uncertainty on the LR distribution. There are other sources of uncertainty which we do not investigate in this work but should still be recognised, such as the number of contributors, the biological models underpinning the statistic and the potential for errors in the generation of the observed data.

The uncertainty in selecting an appropriate value of F_{st} has long been recognised. It is the authors' experience that a commonly used approach to this has been to assign a value believed to be at the top end of the plausible range.

Another matter of significant importance is the presence of relatives as potential alternative donor(s) of the DNA. [17] Traditionally, this matter is subsumed in the formation of the propositions being considered e.g.: *The DNA profile has originated from an unrelated individual from a certain population*. A less common approach is to produce two (or more) LRs; one considering the proposition that the donor is unrelated to the person of interest (POI) and one for the proposition that the donor is a relative of the POI.

A potential solution has been known for some time and is termed the 'unifying formula' [18–20]:

$$LR = \frac{Pr(O|H_1)}{\sum_i Pr(O|H_i)Pr(H_i|H_d)} \quad (1)$$

where typically i is the ith person (related or unrelated) under consideration. H_i is then the proposition that the ith person is the source of the DNA and H_1 is the proposition that the POI is the source of the DNA. In practice, this unifying formula cannot be implemented as when taken to its extreme a different hypothesis is generated for everyone on Earth. Hence i ranges from zero up to the size of the global population, i.e., encompassing every individual in every population. A plausible simplification is to change the meaning of i, firstly to be considered within one population at a time and secondly to be a relationship group, for example individuals whose relationship to the POI is unrelated, parent, child, sibling, cousin, etc. We term this method the 'unified method'. Making the described simplification, the prior may be assigned as the probability of someone in the population being related to the POI with relationship type i. These proportions can be reasonably estimated by making assumptions about population and family structure (see appendix A). Alternatively there may be additional evidence that informs the prior that, say, a brother is the donor of the profile.

A second method that could be used to account for population composition is to generate LRs for each relationship in the proportions that those relationships exist within the population, as estimated from available census data. We term this second method the 'picking method'. The picking method will likely produce an LR distribution over a wider range of values than the unifying method, with a heavy left tail attributed to relatives.

Whilst much is known about the effect of allele frequency variation on the LR distribution the combined effect of these additional sources of uncertainty has not previously been investigated. In this paper we report the effect of each of these sources of uncertainty on a likelihood ratio (LR) calculation and demonstrate how these sources would affect a reported value. The incorporation of these factors into the LR allows the scientist to report a CI using statements such as "*I am 99% sure that the true LR is greater than X*", or "*the LR is above X with 99% probability*". It also removes the need to stipulate that the alternative donor is unrelated when forming the propositions.

Statisticians, and other scientists, commonly use sensitivity analyses to understand how the behaviour of a system, or a model, changes with respect to changes in the inputs. In the forensic DNA interpretation context, parameters are varied over a plausible range and a number of LRs are produced. A statement is then made based on these different numbers. The similarity of this approach to the one described, we hope, will become obvious.

2. Method

A single source AmpF*l*STR® ProfilerPlus™ profile with an unambiguous genotype ($w_i = 1$) at all loci except one was artificially created to minimise the number of variables. The one ambiguous locus had a single peak at a height where dropout, although unlikely, was permitted. At this the locus the designation was *a, Q* where the *Q* allele could be any allele other than *a*.

Empirical distributions of the logarithm base 10 of the LR, $\log_{10}$(LR), distributions were generated from 10,000 HPD iterations, drawing from prior distributions of the parameter of interest in each analysis. The $\log_{10}$(LR) distributions are used for plotting Figures. LR values are calculated using the sub-population model of [21] using propositions:

H_1: The POI is the source of the DNA.

H_2: Someone other than the POI is the source of the DNA.

2.1. Considering only one factor at a time

Throughout this paper we explore the effects of each factor of variation on the LR, holding all other factors constant. To hold allele frequencies constant (in the assessment of all factors of uncertainty other than allele frequency) the database size was inflated to 1 billion (again artificially in the calculation) to effectively remove allele frequency variation. To hold F_{st} constant we use point values in LR calculations that correspond to the mean of the Beta distribution being used to model F_{st} distribution. To hold weights constant we use point values of 0.995 and 0.005 for the locus which is not completely resolved. In the case of population composition, when we wish to remove the effect of population composition on the LR we consider all members of the population as unrelated.

3. Results

3.1. Allele frequencies

Allele probabilities were modelled with a Dirichlet distribution [14] and evaluated by resampling allele counts from gamma distributions and renormalising at each iteration as described in Taylor et al. [3]. Allele frequency estimates from a pan-Australian Caucasian database [22] were used for all simulations. The size of the database was artificially changed to the values of 50, 100, 500, 1000 and 10,000 people and the LR distribution plotted for each database size Fig. 1.

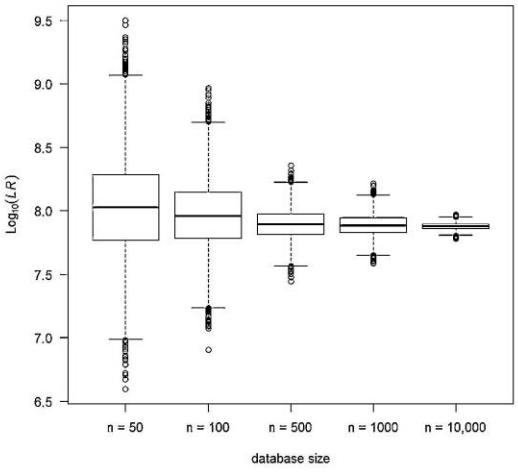


Fig. 1. $\log_{10}$(LR) distributions calculated using differing database sizes.

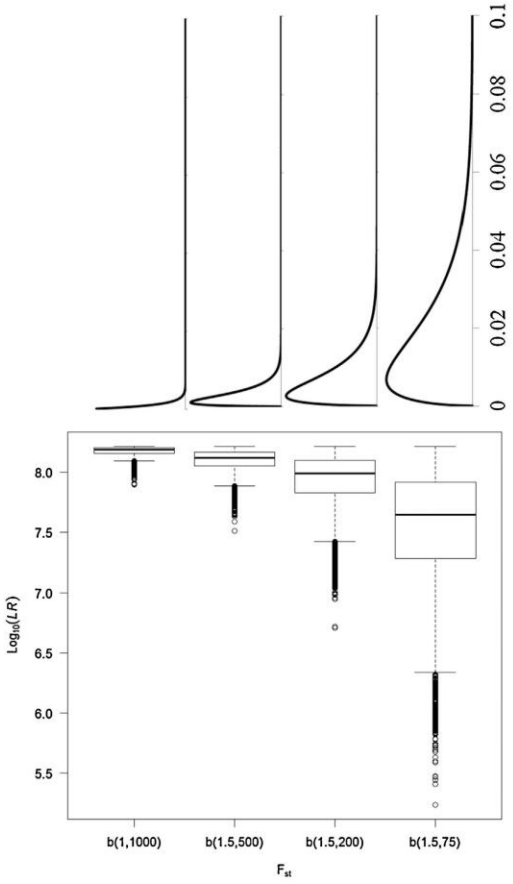


Fig. 2. $\log_{10}$(LR) distributions using four different F_{st} distributions (shown above the boxplot categories).

3.2. F_{st}

The probability that a pair of alleles one taken from each of two people in the population are identical by descent, IBD, will be different for each pair of two individuals. Hence F_{st} within a population may be usefully thought of as having a distribution. This distributions is likely to be asymmetrical, positively skewed, and is constrained to lie between zero and one. The Beta family of distributions have these properties, and so were selected to model the behaviour of F_{st}. Typical values of F_{st} used for forensic LR calculations tend to range between 0 and 5% [23]. We test a number of F_{st} distributions representing differing levels of co-ancestry. Fig. 2 shows the distribution of LRs and the F_{st} distribution used in the calculation.

3.3. Genotype set weights

The weights, w_i, for the genotype set combinations, S_j, were calculated using MCMC as described in Taylor et al. [3]. These w_i values correspond to counts of MCMC iterations where the specified genotype set has been the focus of the MCMC. If the

analysis was repeated, then these counts would be slightly different but would make up approximately the same proportion of total iterations. The variability in counts that make up the weights can be modelled using a gamma distribution as they are likely to be asymmetrical and the counts exist over the range (O, X) where X is the number of iterations in the MCMC. A low weight indicates that the specified genotype was the focus of the MCMC less often (attracted a low proportion of the total iterations) and therefore the relative run-to-run variation is expected to be greater.

It is recognised that successive MCMC iterations are not independent samples from the posterior. The effective sample size (corresponding to the number of independent MCMC iterations) was calculated using the package coda [24] in R [25]. The effective sample size (ESS) was then used to generate effective counts (EC) from the genotype set weights. The EC values were drawn from a gamma distribution $\Gamma(EC, 1)$ and normalised back into weights for use in the LR in each Monte Carlo iteration of the HPD calculation.

The variation in w_i will not only depend on the size of w_i but also on the total number of iterations the MCMC has been allowed to run. An obvious solution to difficulties caused by weight variation is to run each MCMC analysis for an extended number of iterations, making the EC so large that the MCMC run-to-run variation becomes insignificant. This approach may be impractical within a case-working forensic laboratory, where limited computation power may be expected and rapid case turnaround times are required. We test a number of weights with a fixed ESS of 100,000 (Fig. 3) and a number of ESS values with a fixed weight of 0.01 (Fig. 4).

3.4. Fraction of population related to the person of interest

The fraction of population related to the POI was taken into account using two different methods. We refer to the first method as the 'picking method', and the second as the 'unified method' [26]. The fraction of total population related to the POI was assigned using a number of simplifying assumptions about population behaviour that allow numbers of relatives of the POI in a population of size P to be directly related to the average number of children (n) had by couples within the population. Two populations were created, one with a high fraction of relatives and one with a low fraction. We place the details of this approximation

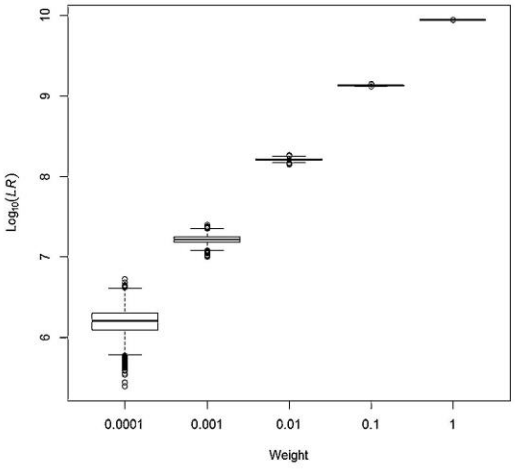


Fig. 3. $\text{Log}_{10}(\text{LR})$ distributions using ESS of 100,000 and different weights.

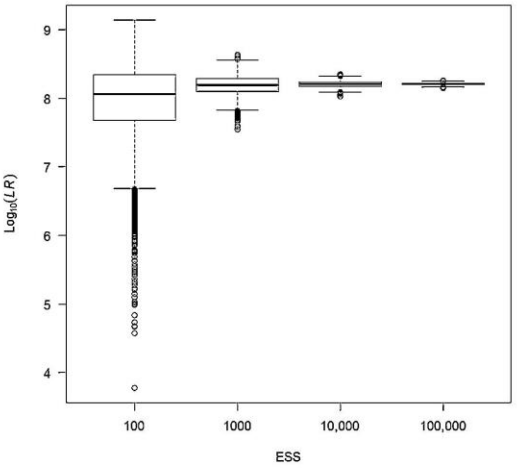


Fig. 4. $\text{Log}_{10}(\text{LR})$ distributions using a weight of 0.01 and differing ESS values.

in Appendix A. These fractions are used as the priors in Eq. (1) which models the situation of no additional information pointing towards or away from any particular relative.

3.5. Picking method

Using the proportion of the total population that the specified relationship comprises, an LR was calculated which considers either a related or unrelated person to the POI in H_i i.e:

$$\text{LR}_i = \frac{\Pr(O|H_1)}{\Pr(O|H_i)} \tag{2}$$

For I Monte Carlo iterations there will be $I \times r_i$ LRs considering a person as the source of DNA in H_i who has a relationship i to the POI. Where r_i is the proportion of the population with relationship i. Note that in Eq. (2) the LR is considering only one type of relationship of the true donor to the POI in the denominator, whereas in Eq. (1) the LR stratified the denominator across all possible relationships.

The picking method represents a discreet equivalent of the continuous methods of accounting for uncertainty outlined in the paper, i.e. a relationship type is drawn from a discreet prior distribution of relationship types and then the LR is calculated given that chosen relationship. To the authors' knowledge this method of accounting for a population's relatedness to a POI has not previously been explored.

3.6. Unified method

Again using the proportion of the total population that the specified relationship comprises, an LR was calculated weighted by the relationship types.

$$\text{LR} = \left(\sum_i \frac{r_i}{\text{LR}_i}\right)^{-1}$$

Where LR_i is the LR considering a relative of type i as the source of DNA in H_2.

Fig. 5 shows two LR distributions for two different population structures. The first represents a large population size (1 million) where the number of children per family is low (2) and the second

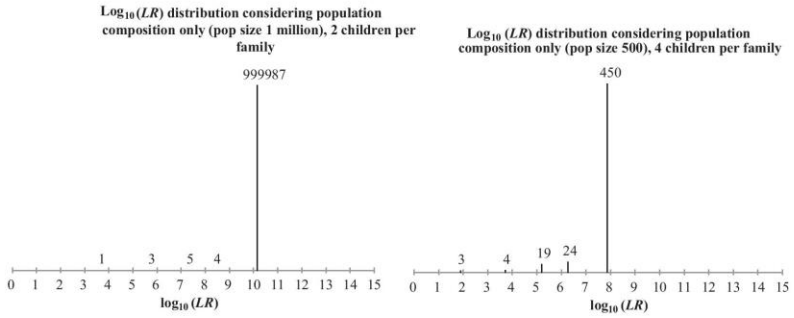


Fig. 5. Log_{10}(LR) distributions produced considering related or unrelated individuals (to the POI) as the source of DNA for two different population compositions.

represents a small population size (500) where the number of children per family is much higher (4). In fig. 5 the position of the vertical lines on the *x*-axis represents several LR values that can be obtained by considering someone other than the POI as the source of the DNA (related or unrelated). The height of the vertical lines represents the proportion of the population that would generate that LR, with the exact number given.

3.7. All uncertainty

Two extremes were considered for each parameter. These extremes represented a low and a high variance for that parameter, so that the range of effects on the LR distribution could be observed. Two ethnic groups were chosen to use as examples: Australian Caucasian [22] and Australian Aboriginal [27]. Assigned F_{st} values for these populations are 0.01 and 0.05 respectively [28–31], representing their markedly different levels of co-ancestry. A Beta distribution was fitted to the F_{st} values using maximum likelihood estimation. The estimated parameters for the Beta distributions are: Beta(0.98, 567.2) for the Caucasian population (low variation scenario), and Beta(0.3, 32.7) for the Aboriginal population (high variation scenario). Using these values, the associated distributions have respective means at 0.0017 and 0.0090, which are markedly lower than the conservative values of 0.01 and 0.05 typically in use.

The low variation scenario had the following properties:

- A large database size (N = 18,116 individuals [22]).
- A low level of co-ancestry where the range of plausible F_{st} values is low (equivalent to an Australian Caucasian population).
- A POI reference that matched at all unambiguous loci and was *homozygous* at the ambiguous locus.
- A large population size with few relatives in it.

The high variation scenario had the following properties:

- A small database size (N = 50, using frequencies from [22] but with N artificially changed in the calculation).
- A higher level of co-ancestry where the range of plausible F_{st} values is high (equivalent to an Australian Aboriginal population).
- A POI reference that matched at all unambiguous loci and was *heterozygous* at the ambiguous locus.
- A small population size with many relatives in it.

The distribution of LRs for each scenario is shown in Fig. 6.

Fig. 7 shows the same scenarios as seen in Fig. 6, but using the unifying method for familial relationships.

3.8. Example

To demonstrate the effect of accounting for multiple sources of uncertainty within the LR, an artificially constructed four person mixture was analysed under two scenarios; firstly with only sampling variation taken into account and then with all variation sources included. The artificial profile had one major and three roughly equal minor contributors. In this example, a POI reference profile corresponding to one of the minor contributors was compared to the mixed profile to generate a LR distribution. The LR distribution generated under the first scenario (as per current

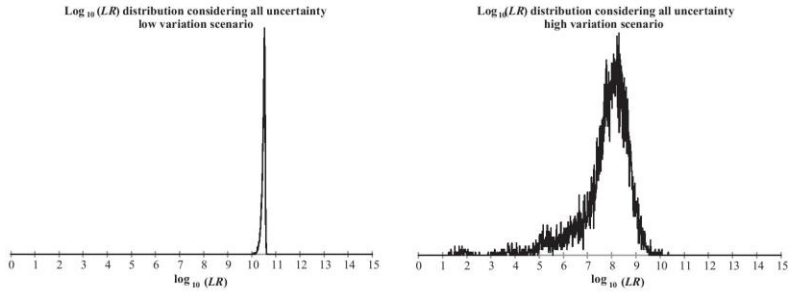


Fig. 6. Log_{10}(LR) distributions of the low and high variation scenarios using the picking method for familial relationships.

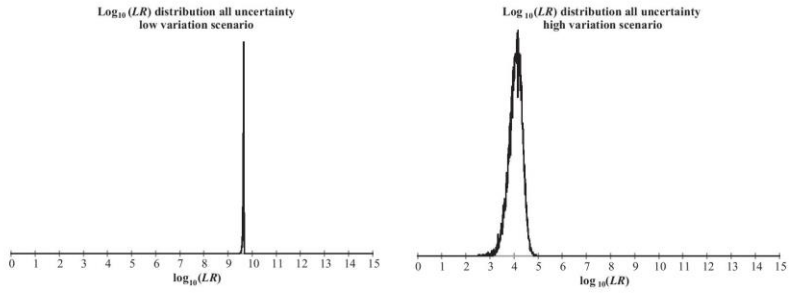


Fig. 7. Log_{10}(LR) distributions of the low and high variation scenarios using the unifying method for familial relationships.

forensic practice) was compared to the distribution created under the second scenario. The following properties were chosen to reflect a real casework situation:

- A South Australian Aboriginal allele frequency database, $N = 325$ individuals [27]
- An F_{st} distribution of Beta(0.3, 32.7) to correspond to an Australian Aboriginal population
- A population size of $P = 26{,}000$ [32] with $n = 3$ children per couple
- The unifying method to account for relatives of the POI in the population
- H_1: The mixture represents DNA from the POI and three unknown individuals, H_2: The mixture represents DNA from four unknown individuals
- An MCMC effective sample size of 772

The comparison of the two analyses can be seen in Fig. 8.

Inspection of Fig. 8B shows that including all uncertainty in the estimation of the LR pushes the lower end of the distribution below zero, into the range where the alternate hypothesis is supported. Note that the effect of the F_{st} on the LR distribution is more prominent given the accumulative effect that F_{st} has across genotype probabilities that contain three or four unknowns. The result is that whilst the LR distribution in Fig. 8B is more negatively skewed, its mode is pushed to the right of where it was in 8 A. Note that a graph similar to Fig. 8B has not been produced using the picking method to account for population relatedness to the POI. The reason for this is that given the population size and average number of children per family, the number relatives of the POI expected in the population is small, and there would be no visual (and very little numerical) difference to the Figure already produced. The reported LR taking into account all uncertainty parameters (Fig. 8B) will be either above or below 1 depending on whether the point estimate or a CI is chosen. The exact reportable values for the results seen in Figs. 6, 7 and 8 can be seen below in Table 1.

4. Discussion and conclusions

Our work shows that any of the sources of variation outlined in this study could dominate the total variation in the LR distribution under the right circumstances. For example, uncertainty in genotype weight distributions can dominate the LR variation if one or more weights for relevant genotype sets are low, or are in combination with inadequate MCMC iterations and the calculation performed uses a large population database. However if the weights are high for relevant loci and the MCMC is run for many iterations, but a small population database is used then the allele frequency distribution will dominate as the LR variation. In most cases some combination within the ranges tested in this study will be present. In combination, all factors of uncertainty will interact to produce an LR distribution that is wider than if any individual component is considered in isolation.

Although many sources of variation will exhibit interactions, some interactions of interest are worth mentioning with regards to F_{st}. Consideration of the Balding and Nichols [21] sub-population formula suggests the F_{st} variation will have a more pronounced effect on the LR distribution as more unknowns are considered in the genotype probabilities. It is also expected that F_{st} distribution will have a larger effect on *LR* distribution when considering rare genotypes due to the application of F_{st} in the sub-population

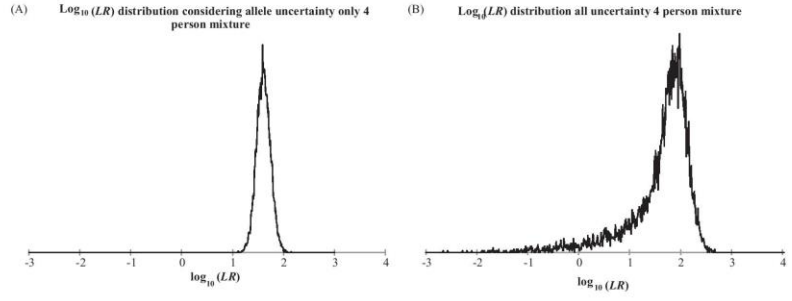


Fig. 8. Log_{10}(LR) distributions accounting for (A) allele frequency variation only (using mean F_{st} value) and (B) all variation parameters.

Table 1
Reportable LR values showing point estimate, 99% 1-sided credible interval (CI) considering only allele frequency variation (using mean F_{st} values) and the same CI considering all variation parameters.

Single source profile	Low variation scenario	High variation scenario
Point estimate	3.03×10^{10}	7.79×10^{7}
CI allele frequency variation only	2.70×10^{10}	1.47×10^{7}
CI all uncertainty (picking method)	1.70×10^{10}	5.13×10^{3}
CI all uncertainty (unifying method)	3.69×10^{9}	2.00×10^{3}
Four person mixture		Artificial crime sample
Point estimate		28.69
CI allele frequency variation only		19.95
CI all uncertainty (unifying method)		0.19

formula. This has the consequence that the LR distribution resulting from small database size can be partially masked by a wide F_{st} distribution.

Either method of accounting for population composition gives comparable results when considering a 99%, one-sided CI. However the picking method has several disadvantages. When population size is large in comparison to the number of relatives expected (so that the chance of picking a relative is less than the inverse of the number of HPD iterations) then often the picking method will not ever choose a relative and so the population composition has no effect on the LR distribution. In addition, there is no easy way to generate an LR point estimate using the picking method as the distributions can be multimodal.

It is typical in forensic laboratories to report an LR point estimate and a credible interval that takes allele frequency variation into account. Carrying out such a calculation can give the false impression that all uncertainty has been taken account of, and this may invite false statements or conclusions. In reality allele frequency variation is only one source of variation, and may not even be the greatest source of total variation within the LR. A reported LR probability interval which takes all sources of uncertainty into account can be similar in magnitude to the allele-frequency-only-LR if the other sources of variation are low (an order of magnitude as seen in the low variation scenario in Table 1). However, if other sources of variation are high, then the difference between the reported LR when taking only allele frequency and all sources of variation into account can be significant. In the high variation scenario this difference was four orders of magnitude. As seen in the four person example this difference amounted to a reduction in the reported LR of over two orders of magnitude, but perhaps more interestingly in this example, the 99% CI and the point estimate were on opposite sides of the neutrality line, LR = 1.

The use of MCMC methods to determine weightings, and Monte Carlo resampling to carry out the HPD calculations provide a powerful and flexible tool for assessing sources of uncertainty. The risk of using such a system is that the mathematics and general concepts are less transparent to an average user. Detailed training in the concepts of MCMC and HPD are required to avoid a 'black box' system. In addition, the transition to an MCMC based system requires a more complex plan for validation than simpler methods, which previously could be assessed with comparison to relatively simple hand calculations in many cases. Despite these complexities, we feel that the system described within this paper, coupled with appropriate training, would provide a powerful addition to a forensic laboratory's ability to interrogate EPG data.

The question remains, given that we can evaluate all these uncertainties within the *LR* calculations, what value should be reported to a courtroom? There are advocates that argue the most relevant figure to provide is the point estimate [6]. However, in our experience the court often concerns itself with exploring potential sources of uncertainty. If the decision is made to report a CI (or at least to calculate one for reference if required) then it would be difficult to justify only considering one aspect of total variation (i.e. allele frequency estimate, as is currently prevalent practice) and not others if an available method exists to reasonably do so. It is the personal view of the authors that a statistical evaluation of evidence should concede reasonable doubt and uncertainty to the defendant and so a reported value should be some lower quantile (e.g. 95 or 99%) of the LR distribution in criminal matters, but we respect alternative decisions. For civil matters there is often no side to which it is obvious to concede doubt and so a point estimate (mode or 50% quantile) may be the more appropriate figure to report.

Acknowledgements

This work was supported in part by grant 2011-DN-BX-K541 from the US National Institute of Justice. Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the U.S. Department of Justice. We would also like to thank two anonymous reviewers for their helpful comments that improved this paper.

Appendix A

Proportions used for fraction of population related to POI.

The simplifying assumptions used to generate the relatives proportions are:

- There are three generations persisting in the population present in equal proportions.
- Generation 1 is the youngest and has no children.
- Generation 2 is the middle and has both children and parents surviving.
- Generation 3 is the oldest and has no surviving parents.
- There is only one union between two families i.e., there are no instances of two siblings from one family bearing children with two siblings from another family.
- Couples form from within their own generation.
- There are no instances of inbreeding closer than or equal to the first cousin level.
- Each family has the same structure and number of children.
- There are no instances of early death.

Table 2
Fraction of individuals in a population with known relationship.

Relationship type, i		The fraction of the population with relationship i, r_i	
		Low relatedness population	High relatedness population
Unrelated	1	0.999987	0.90000
Siblings	2	0.000001	0.00600
Children	3	1.33E-06	0.00600
Parents	4	1.33E-06	0.00533
Uncle/auntie	5	1.33E-06	0.00267
Niece/nephew	6	1.33E-06	0.00800
Grandparents	7	1.33E-06	0.01600
Grandchildren	8	1.33E-06	0.00267
First cousins	9	0.000004	0.04800

Using these assumptions the number of individuals of each relationship type to the POI were calculated for two populations; a large population with a low number of children per family and a small population with high numbers of children per family as seen in Table 2.

References

[1] P. Gill, C.H. Brenner, J.S. Buckleton, A. Carracedo, M. Krawczak, W.R. Mayr, et al., DNA commission of the International Society of Forensic Genetics: Recommendations on the interpretation of mixtures, Forensic Sci. Int. 160 (2006) 90–101.
[2] T.M. Clayton, J.S. Buckleton, Mixtures. Forensic DNA Evidence Interpretation, CRC Press, Boca Raton, 2004, pp. 217–227.
[3] D. Taylor, J.-A. Bright, J.S. Buckleton, The interpretation of single source and mixed DNA profiles, Forensic Sci.e Int.: Gen. 7 (2013) 516–528.
[4] M.W. Perlin, M.M. Legler, C.E. Spencer, J.L. Smith, W.P. Allan, J.L. Belrose, et al., Validating TrueAllele® DNA Mixture Interpretation, J. Forensic Sci. 56 (2011) 1430–1447.
[5] J.M. Curran, A MCMC method for resolving two person mixtures, Sci. Justice 48 (2008) 168–177.
[6] C.H. Brenner, DNA frequency uncertainty – why bother (1997)
[7] P.I. Good, Applying Statistics in the Courtroom, Chapman&Hall/CRC, London, 2001.
[8] I.W. Evett, Weir BS. Interpreting DNA Evidence – Statistical Genetics for Forensic Scientists, Sinauer Associates Inc., Sunderland, 1998.
[9] R. Chakraborty, M.R. Srinivasan, S.F. Daiger, Evaluation of standard errors and confidence intervals of estimated multilocus genotype probabilities and their implications in DNA, Am. J. Hum. Genet. 52 (1993) 60–70.
[10] NRCII, National Research Council Committee on DNA Forensic Science, The Evaluation of Forensic DNA Evidence, National Academy Press, Washington, D.C, 1996.
[11] R. Chakraborty, Sample size requirements for addressing the population genetic issues of forensic use of DNA typing, Hum. Biol. 64 (1992) 141–160.
[12] B. Budowle, K.L. Monson, R. Chakraborty, Estimating minimum allele frequencies for DNA profile frequency estimates for PCR-based loci, Int. J. Legal Med. 108 (1996) 173–176.
[13] D.J. Balding, Estimating products in forensic identification using DNA profiles, J. Am. Stat. Assoc. 90 (1995) 839–844.
[14] J.M. Curran, J.S. Buckleton, C.M. Triggs, B.S. Weir, Assessing uncertainty in DNA evidence caused by sampling effects, Sci. Justice 42 (2002) 29–37.
[15] C.M. Triggs, J.M. Curran, The sensitivity of the Bayesian HPD method to the choice of prior, Sci. Justice 46 (2006) 169–178.
[16] J.M. Curran, J.S. Buckleton, An investigation into the performance of methods for adjusting for sampling uncertainty in DNA likelihood ratio calculations, Forensic Sci. Int.: Gen. 5 (2011) 512–516.
[17] J. Buckleton, Triggs C. Relatedness and DNA: are we taking it seriously enough? Forensic Sci. Int. 152 (2005) 115–119.
[18] D.J. Balding, P. Donnelly, Inference in forensic identification, JRSS A 158 (1995) 21–53.
[19] D.J. Balding, P. Donnelly, Inferring identity from DNA profile evidence, in: Proceedings of the 92 National Academy of Science, USA, (1995), pp. 11741–11745.
[20] J.S. Buckleton, A framework for interpreting evidence, in: J.S. Buckleton, C.M. Triggs, S.J. Walsh (Eds.), Forensic Dna Evidence Interpretation, CRC Press, Boca Raton, Florida, 2005.
[21] D.J. Balding, R.A. Nichols, DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands, Forensic Sci. Int. 64 (1994) 125–140.
[22] S.J. Walsh, J.S. Buckleton, Autosomal microsatellite allele frequencies for a nationwide dataset from the Australian Caucasian sub-population, Forensic Sci. Int 168 (2007) e47–e50.
[23] J.S. Buckleton, C.M. Triggs, S.J. Walsh, DNA Evidence, CRC Press, Boca Raton, Florida, 2004.
[24] M. Plummer, N. Best, K. Cowles, K. Vines, CODA: convergence diagnosis and output analysis for MCMC, R News 6 (2006) 7–11.
[25] M. Plummer, Bayesian graphical models using MCMC, RJAGS (2012).
[26] D.J. Balding, Weight-of-evidence for Forensic DNA Profiles, John Wiley and Sons, Chichester, 2005.
[27] D.A. Taylor, J.M. Henry, Walsh SJ. South Australian Aboriginal sub-population data for the nine AMPFlSTR® Profiler Plus™ short tandem repeat (STR) loci, Forensic Sci. Int.: Gen. 2 (2008) e27–e30.
[28] S.J. Walsh, R.J. Mitchell, J.M. Curran, J.S. Buckleton, The extent of substructure in the indigenous Australian population and its impact on DNA evidence interpretation, Int. Congr. Ser. 1288 (2006) 382–384.
[29] J. Buckleton, S. Walsh, J. Mitchell, Autosomal microsatellite diversity within the Australian population, Report of the National Institute of Forensic Sciences Standing Committee on Sup-Population Data (2007).
[30] S.J. Walsh, J. Buckleton, Autosomal microsatellite allele frequencies for 15 regionally defined Aboriginal Australian population datasets, Forensic Sci. Int. 168 (2007) e29–e42.
[31] S.J. Walsh, R.J. Mitchell, N. Watson, J.S. Buckleton, A comprehensive analysis of microsatellite diversity in Aboriginal Australia, J. Hum. Genet. 52 (2007) 712–728.
[32] Australian Bureau of Statistics. 2011 Census QuickStats: South Australia 2011

3.4: Extending the use of the LR for complex situations

Most ideas develop from simple beginnings. In this vane, the very early versions of STRmix™ had limited functionality compared to the abilities in current versions of the software. Early on the two main functions of the program were to deconvolute a DNA profile into a list of potential contributing genotypes, with associated weights (that represented a goodness-of-fit for each genotype in describing the profile) and secondly to calculate an LR when the deconvolution results were compared to a reference DNA profile. A relatively straight forward extension of this second function was that if an LR can be calculated for the comparison of a reference to a mixed DNA profile, then there was no reason that it couldn't do so for 10, or 100 or 1 million references, one at a time, in an automated fashion. All that was needed was for those reference DNA profiles to be listed in a file somewhere that STRmix™ could access. What has just been described is a variant on a very common practise in forensic biology known as database searching. This is the process of searching a database (local or national) for matching copies of an evidence DNA profile that has been generated from an unsolved crime. Equally, database searching can be carried out by searching the reference of a known individual to determine whether they are associated with any evidence profiles from currently unsolved crimes. The standard practise, pre-STRmix™, was that only DNA profiles originating from a single individual (or a manually interpreted single contributor's profile from a mixture) could be searched against a database in a simple process of matching arrays of numbers. With STRmix™ the additional ability was obtained to assign an LR to each member in the database, considering them as a potential contributor of DNA to an unresolvable DNA mixture. Then, rather than having a match/no-match criteria for identifying people in the database a sliding scale of support for them being a DNA donor is generated.

This opened a vast number of DNA profiles, from unsolved crimes, to database searching that previously could not be used to assist with the solving of the crime. So successful was the idea that the New Zealand forensic laboratory (ESR) had the process programmed into their laboratory information management system so that the searches could be conducted in an automated manner on a day to day basis, without the need for exporting and importing data from STRmix™. ESR has shown great success with this feature. At Forensic Science SA the process of 'mixture searching' was introduced in 2016, which sees unresolvable, mixed DNA profiles able to be searched against the state database. There have been instances of three contributors to a single DNA mixture all being identified in the one search. Anecdotal information also suggests that other Australian state laboratories are using the feature to great success. On a national stage database searching and matching still only occurs using single contributor profiles and a simple comparison of numerical allele values. Perhaps in the future a probabilistic approach could be implemented, which would have enormous benefits.

Questions of performance and reliability arose from using the database searching functionality. One of these was the question of how commonly individuals would yield support for being a contributor of DNA to a DNA profile, when in fact they were not DNA donors. In forensic biology, this is commonly referred to as 'adventitious matching', although the term comes from many years ago when comparison outcomes were more binary (i.e. a profile would either match or not match another) and the term 'matching' doesn't sit so well in a continuous world and a suggestion has been to relabel the phenomenon as obtaining 'misleading LRs'. Regardless, the term is common enough that it is still used today. The first paper in this section describes the functionality of database searching in a continuous manner (such as used by STRmix™) and

partially addresses the issues of adventitious matching. The topic is explored in more depth in chapter 5, where some publications are provided that are specifically focussed on the point.

The second paper in this section demonstrates the mathematics required to consider the support the LR provides to a nominated individual being a contributor of DNA to a mixture, if the alternative that must be considered is that it is one of their relatives (as opposed to a 'random' person from the population). This work came about for two reasons. Firstly, with the advent of DNA profiling kits that tested 20 or more DNA locations the LR being generated, when considering the alternate DNA source as being an unrelated person from the population, were in the range of 10^{20} to 10^{30}. The defence community started to shift the questioning to ask '*what if the person who committed the crime was the sibling/parent/cousin/etc of the accused?*'. There is a common defence stance in forensic biology known as 'the brothers defence' where the defendant is legitimately claiming it is their sibling who is the real perpetrator of the crime. What was being asked in court was not exactly this scenario, as there was no reason for any particular relative to be considered an alternate offender, rather it was a series of what-if questions. What the defence community was actually asking for (although they didn't realise it) was for some way to consider that a proportion of the population would be related to the defendant and any one of them had, *a priori*, an equal probability of being the alternate offender as any one unrelated person. They were referring to the 'unified LR', which was mentioned earlier in this chapter.

As a bi-product of this work all the tools required to carry out another common forensic activity known as familial searching were present. Rather than searching a database for a specific person's profile, familial searching looks for any potential relatives of that person. In a 'unrelated' LR the probability of obtaining the evidence is calculated given the two propositions:

1) The DNA came from the POI and others
2) The DNA came from people other than the POI, all unrelated

And for the relatives LRs the probability of obtaining the evidence is calculated given:

3) The DNA came from the POI and others
4) The DNA came from people other than the POI, one of who was related to the POI

Then a familial search is just the probability of the evidence given propositions 4 and 2 from above. In a similar manner to regular database searching, STRmix™ could then (and eventually was) programmed to carry out familial searches on complex unresolvable mixtures. The mathematics required for the consideration of relatives, and the manner in which it can be used to carry out familial searches is given in the second paper in this chapter.

Due to the complexity and public sensitivities surrounding familial searching it is not carried out with the same regularity as standard database searching. A number of laboratories that do carry out familial searching, do not wish this to be publicly advertised. In South Australia, FSSA has been carrying out familial searches since 2008 and started using STRmix™ to do so in 2013. All familial searches have only been carried out on single sourced evidence profiles, and one of the searches conducted at FSSA has led to an arrest in SA. The only exception to

this is one familial search carried out in SA on a two-person mixed DNA profile from an old bone sample of an unidentified deceased child, in the hope that their relative may be in the database. The search itself went fine, but did not identify a relative. A relative of the child was eventually found, although not through familial searching, but rather good police work.

Manuscript: Searching mixed DNA profiles directly against profile databases. JA Bright, D Taylor, J Curran, J Buckleton. (2014) Forensic Science International: Genetics 9, 102-110 – *Cited 18 times*

Statement of novelty: Until this point it was generally believed that only single source profiles could effectively be searched against a database. This work presents the means by which STRmix™ can be used to compare complex and unresolvable mixtures to a database of potential contributors.

My contribution: I was a part contributor to the theory and contributor to the writing of the manuscript.

Research Design / Data Collection / Writing and Editing = 20% / 0% / 10%

Additional comments: While not a main contributor of work to this paper, I have included it in my thesis for a few reasons:

1) The mathematics for calculating the LR for the comparison of a reference to a complex mixture had not routinely been expanded to do this en masse for an array of references in a database. This feature was programmed by me and implemented in STRmix™ software in 2012. The paper below (while I was not dominant in its construction), used this feature and is the only published demonstration of the power of the technique
2) The paper highlights one of the important points of impact that STRmix™ has provided to the forensic community. Mixture searching is now routinely carried out by many forensic laboratories and has resulted in numerous examples of intelligence being provided to police that would have otherwise been missed. I go into this in more detail in chapter 9. I feel it is therefore important to include this work in the thesis as the published demonstration of the technique.

Forensic Science International: Genetics 9 (2014) 102–110

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

Searching mixed DNA profiles directly against profile databases

Jo-Anne Bright [a,b,*], Duncan Taylor [c], James Curran [b], John Buckleton [a]

[a] Institute of Environmental Science and Research Limited, Private Bag 92021, Auckland 1142, New Zealand
[b] Department of Statistics, University of Auckland, Private Bag 92019, Auckland 1142, New Zealand
[c] Forensic Science South Australia, 21 Divett Place, Adelaide, SA 5000, Australia

ARTICLE INFO

Article history:
Received 2 August 2013
Received in revised form 1 December 2013
Accepted 3 December 2013

Keywords:
Forensic DNA
Database
Continuous models
Likelihood ratio

ABSTRACT

DNA databases have revolutionised forensic science. They are a powerful investigative tool as they have the potential to identify persons of interest in criminal investigations. Routinely, a DNA profile generated from a crime sample could only be searched for in a database of individuals if the stain was from single contributor (single source) or if a contributor could unambiguously be determined from a mixed DNA profile. This meant that a significant number of samples were unsuitable for database searching. The advent of continuous methods for the interpretation of DNA profiles offers an advanced way to draw inferential power from the considerable investment made in DNA databases. Using these methods, each profile on the database may be considered a possible contributor to a mixture and a likelihood ratio (*LR*) can be formed. Those profiles which produce a sufficiently large *LR* can serve as an investigative lead.

In this paper empirical studies are described to determine what constitutes a large *LR*. We investigate the effect on a database search of complex mixed DNA profiles with contributors in equal proportions with dropout as a consideration, and also the effect of an incorrect assignment of the number of contributors to a profile. In addition, we give, as a demonstration of the method, the results using two crime samples that were previously unsuitable for database comparison. We show that effective management of the selection of samples for searching and the interpretation of the output can be highly informative.

1. Introduction

DNA databases can be powerful tools in the identification of individuals of interest during a criminal investigation. Typically, DNA databases consist of two sub databases; one containing profiles from known individuals who have either volunteered or been compelled to provide a sample (the database) and the other is a database of profiles collected from samples associated with crime scenes [1] (the crime sample database). The records in the separate databases can be compared to each other to link individuals with crime scenes. This comparison process typically takes a crime scene profile and compares it to each database sample in turn. Often a count is made of concordant and non-concordant alleles. A wild card designation may be included in the crime sample profile or more rarely in the database profile. The wildcard is deemed to be concordant with any allele. Most search algorithms are set up to compare two alleles per locus from the crime sample profile with the two alleles per locus from the database profiles. This approach restricts profiles suitable for searching to single source profiles or a single source component inferred, either completely or partially, from a mixed DNA profile. Attempts to extend the utility of databases have included searching against a reduced list of genotypes created from a mixture [2].

This investigative intelligence is provided to investigators to assess in conjunction with the wider case information.

If both the crime sample profile and the database profile are full multilocus profiles then the chance of an adventitious match is small. Adventitious matches are more likely with low level, partial or mixed profiles. Many databases now include profiles from superseded multiplexes which may have as few as six loci scored. Adventitious matches, although expected, can reduce the credibility of the databank operation or even the forensic use of DNA. As an example, the discovery of a number of partial matches in the Arizona database led to considerable discussion including some adverse comment even though these partial matches occurred at approximately the expected rate (see Mueller [3])

The quality of the database can be ensured by legislation such as restricting the type of sample, setting a minimum number of alleles required for database entry, by mandatory participation in quality assurance programmes (as in the USA), and by participation in external audits and proficiency tests [4].

* Corresponding author at: Institute of Environmental Science and Research Limited, Private Bag 92021, Auckland 1142, New Zealand. Tel.: +64 9 815 3940; fax: +64 9 849 6046.
E-mail address: Jo.bright@esr.cri.nz (J.-A. Bright).

1872-4973/$ – see front matter © 2013 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.fsigen.2013.12.001

Whilst it is relatively easy to meet a very high standard for reference or individual profiles, the profiles from crime scenes are frequently compromised in quality.

The likelihood ratio (*LR*) is generally accepted as the most powerful and relevant statistic for the calculation of the weight of the DNA evidence [5]. It is the ratio of the probability of the observed crime stain (O) given each of two competing hypotheses, H_1 and H_2, and given all the available information, I. Mathematically, we express this as:

$$LR = \frac{\Pr(O|H_1, I)}{\Pr(O|H_2, I)}$$

Typically database search algorithms do not calculate an *LR* but simply report the number of concordant and non-concordant alleles. However *LR* based approaches have proven useful in familial searching [6,7]. For unresolvable or low level mixtures the use of an *LR* confers considerable advantages as we demonstrate in this paper.

Stochastic events such as heterozygote imbalance, allelic dropout, locus dropout, and allelic drop in can complicate interpretation [8–10]. The uncertainty in the numbers of contributors and stutter, a by-product of the PCR process, can further complicate profile interpretation whenever stutter peaks are of a similar height to the minor allelic peaks in mixed DNA profiles.

The number of contributors to a mixed DNA profile is never known with absolute certainty. It may be easily determined if the number of alleles is known. It is the step of inferring how many alleles are present from the peaks that is the source of uncertainty. Some peaks are not allelic (for example artefacts or stutter peaks) and some represent contributions from two or more alleles from the same or different individuals superimposed. Some alleles may not have produced a peak due to dropout. At high sensitivity it is possible that some peaks are formed by alleles from the laboratory environment, termed drop-in. Information from replicate amplifications and in certain situations Y STR analysis can be helpful in providing a reasonable estimate of the number of contributors. Statistical methods such as maximum likelihood [11] or Bayesian networks [12] are more statistically sound, and can compensate for artefacts such as stutter, and dropout.

The suggestion that there is a correct number of contributors for every profile would seem self-evidently true but overlooks the fact that this number is inherently unknown and that it is conditioned on what is known about the profile. It should be noted that there is no reason for the number of contributors to be the same under the hypotheses H_1 and H_2. However, proposing an unreasonable number of contributors under the defence hypothesis and holding the number under the prosecution hypothesis at a reasonable assignment will increase the *LR*, favouring the prosecution hypothesis [13].

Complications in profile interpretation have led to a recent push for forensic laboratories to introduce improved models for DNA interpretation. This is motivated by the difficulties traditional methods have with the interpretation of complex profiles [14,15]. The traditional methods of interpretation are described as binary which describes the fact that the probability of the genotype combination under consideration is assigned as zero or one (hence binary) [16]. Following Kelly et al. [17] we denote the genotype of the observed crime stain as O, and the genotypes of proposed donors as G_i for donor i. For an N donor mixture there are N proposed genotypes, G_i for each proposed combination. The jth set of N genotypes is denoted S_j. Binary models assign the values zero or one to the unknown probability $Pr(O|S_j)$ based on heuristics such as heterozygote balance and mixture proportion, the reasonable values of which are informed by empirical data. Essentially, $Pr(O|S_j)$ is assigned a value of zero if the genotype combination falls outside of these heuristics. $Pr(O|S_j)$ is assigned a value of one if it falls within. These binary methods are slowly being replaced by more advanced interpretation methods, such as the semi-continuous models likeLTD and LRmix and continuous models which can take into account stochastic events. STRmix [18,19], TrueAllele [20] and the model described by Puch-Solis et al. [21], are examples of software that employ a continuous model for DNA profile interpretation.

A continuous model uses the quantitative information from an electropherogram (epg) to calculate the probability of the peak heights given all possible genotype combinations, assigning a value or weight (w_i) to the normalised probability $Pr(O|S_j)$. Continuous models can remove some of the qualitative thresholds such as heterozygote balance and may remove some of the subjective decisions required within a binary model. A discussion of the merits of the different interpretation models can be found in Kelly et al. [17].

STRmix assigns a relative weighting to the probability of the epg given each possible genotype combination at a locus. The weights across all combinations at that locus sum to one. Therefore, a single unambiguous genotype combination at any locus would be assigned a weighting of one.

Good quality single source DNA profiles, where stochastic effects are not an issue, are likely to result in a profile of sufficient quality for entry to a crime sample database regardless of the interpretation method used. However mixed profiles, or single source profiles subject to stochastic effects, may not result in a profile suitable for entry to a database using traditional binary methods. Interpretation of these profiles using a continuous model may result in improved profile information and therefore permit database entry. Unless the weight for any given genotype combination is one, assessing the 'quality' of a profile for its suitability for comparison to a database is not straightforward. A guideline for database entry based on some assessment of the risks of loading an incorrectly inferred profile may be employed where the genotype combination of a contributor is ambiguous, such as $w_i > 0.99$. If an individual's profile cannot be reasonably inferred from a DNA mixture, regardless of the interpretation method, then it is unsuitable for entry to a database using traditional database methods.

The number of contributors to a mixed DNA profile (N) cannot be known with certainty. It may be the case that the same electropherogram can be interpreted as having come from several different numbers of contributors. Assigning the probable numbers of contributors to a mixed DNA profile is more complicated with low level profiles. Uncertainty is increased when peaks are close to the limit of detection or there are additional peaks just below the analytical threshold. These cases might invoke the addition of a contributor to a profile. Overestimating the number of contributors to a profile has the potential to generate an *LR* that favours inclusion of known non-contributors, whereas underestimating the number of contributors has the potential to generate an *LR* that favours exclusion of a known contributor. Neither of these outcomes is desirable.

The number of contributors must be specified when using current likelihood ratio implementations for profile interpretation [22]. Direct comparison of mixed DNA profiles, where there are multiple possible genotype combinations at one or more loci, to profiles of individuals within a database, can be undertaken using the output of a continuous method of interpretation with a modified search algorithm using a likelihood ratio framework.

In this paper, a method for database entry and comparison to the New Zealand DNA Profile Databank (DPD)[1] of previously

[1] The NZ DPD was established in 1996 [23] and comprises DNA profiles amplified using the Second Generation Multiplex (SGM, Forensic Science Services, UK), and Applied Biosystems' SGMPlus™ and Identifiler™ multiplexes (Life Technologies, Carlsbad CA). As at April 2013 the DPD comprised 8860 SGM profiles, 65,568 SGMPlus™ profiles and 69,543 Identifiler™ profiles.

unsuitable mixed DNA profiles is described. We examine the efficacy of the method using artificially prepared low level, mixed DNA profiles where the individual contributor profiles are known. We also report the results of two case examples.

2. Method

Database profiles were blood samples or saliva stains on FTA® Classic or Elute card (Whatman, Maidstone, England). The method for processing is described in Bright et al. [24].

Eight artificial mixed DNA profiles were prepared by amplifying extracted DNA from three known sources with the approximate mixture proportions of 10:5:1 (referred to as major:minor:trace) in varying contributor orders. DNA from the eight prepared mixtures and two case examples was extracted using Promega's DNA IQ™ magnetic bead extraction chemistry (Madison, WI) and quantified using Applied Biosystems Quantifiler™ real time PCR quantitation kit (Life Technologies). A target of 1.5 ng of DNA was amplified using Applied Biosystems' Identifiler™ multiplex (Life Technologies, Carlsbad CA) on an Applied Biosystems 9700 thermal cycler with a silver block as per manufacturer's recommendations [25]. Amplified products were separated on an Applied Biosystems' 3130*xl* Genetic Analyser and data was analysed using Applied Biosystems' GeneMapper™ *ID* version 3.2.1 using a 50 relative fluorescent unit (rfu) analytical threshold. Prior to interpretation, the heights of all peaks within the epg of the eight artificial mixtures were halved in order to mimic low level profiles or further modified as described in each experimental method below. Peaks that subsequently fell below 50 rfu were removed prior to interpretation.

In addition, an artificial two person mixture was created by combining the alleles from two known individuals in the proportion 1:1. In one replicate, the peaks were set to a height where dropout would not be a consideration (called 'two person without drop') and in another replicate the peaks were lowered to a height where dropout was very likely ('two person with dropout'). An artificial three person mixed DNA profile was created in a similar fashion with three known DNA profiles in the proportion 1:1:1. Peaks heights were adjusted where dropout was not a consideration ('three person without dropout') and where dropout was expected ('three person with dropout').

All profiles were interpreted using STRmix [26].

2.1. Experiment 1 – testing the effect of the number of contributors in the same mixture proportions

Four artificial mixed DNA profiles (one two- and one three-contributor mixture with and without dropout) were interpreted assuming the known number of contributors. The profiles were compared with 145,470 profiles on the NZ DPD plus the profiles of the known contributors. An *LR* was calculated using the continuous method (LR_C) described in Taylor et al. [18] for each profile from the DPD and the known contributors. Each of the individuals on the database and the known contributors were considered as a potential contributor in turn under the following two hypotheses:

H_1. Database individual and $N-1$ unknown contributors.

H_2. N unknown contributors.

where N is the number of contributors under consideration. As this search is undertaken during the investigative phase, no subpopulation correction was used and the product rule was calculated. An important benefit is reduced computational effort when searching.

A population database comprising allele frequencies of the four major subpopulations within NZ in their approximate proportions as determined in the 2006 NZ Census was used to generate the *LR*.

The *LR* for all known contributors and all adventitious matches (known non-contributors) was recorded.

2.2. Experiment 2 – testing the effect of overestimating the number of contributors

Eight mixed profiles consisting of DNA from known contributors were interpreted as originating from both three (correct) and four (incorrect) contributors. The profiles were compared with 145,470 profiles on the NZ DPD plus the three profiles of the known contributors and an LR_C calculated as described in experiment 1 above. Each of the individuals on the database and the three known contributors were considered as a potential contributor as in experiment 1.

The *LR*s for all known contributors and all adventitious matches (known non-contributors) were recorded. This experiment allows a determination of the *LR*s for the known contributors and known non-contributors. In addition we examine the effect of an incorrect assignment of the number of contributors in the interpretation. It examines the behaviour of the process if the number of contributors is wrongly assessed as one more than the true number.

2.3. Experiment 3 – testing the effect of low level profiles

The eight profiles were further reduced in rfu scale to 10% of the original heights and stochastic effects introduced by the random addition of rfu. This was designed to mimic extremely low level profiles. After reduction in height, all peaks were below 800 rfu with the majority under 400 rfu. All profiles appeared as having only two contributors based on allele count. The profiles were interpreted assuming both two and three contributors. In order to manage run times a 'high risk' database was created by pooling all profiles where the LR_C from experiment 2 was above 100. Within the high risk database were 595 Identifiler, 742 SGMPlus and 92 SGM profiles. The proportions of the different multiplexes within the new high risk profile database were approximately the same as the original database. Each of the individuals on the new high risk database ($N = 1429$) and the three known contributors were considered as a potential contributor and the LR_C calculated as described for experiment 1 above.

2.4. Experiment 4 – testing the effect of underestimating the number of contributors

A four donor profile was artificially constructed from the 50% reduced known three person mixture (Profile 1) used in experiment 2 by adding a fourth contributor in such a way that allele count would not indicate the presence of the fourth contributor. The fourth contributor was added at the same height as the known trace third contributor. Three additional alleles were added to the profile. Where the fourth contributor shared alleles with the known three contributors or had peaks in stutter positions these peak heights were also increased proportionally. Each of the individuals on the original database ($N = 145{,}470$) and the four known contributors were considered as a potential contributors and the *LR* calculated as described for experiment 1 above.

3. Results

3.1. Experiment 1

The artificial two person profile without dropout resulted in an LR_C value above zero for only the two known contributors within the database. There were no adventitious matches to known non-contributors.

The two known contributors to the two person profile with dropout resulted in the highest LR_C values. 2801 individuals within the database also provided adventitious links to this artificial mixed DNA profile, with LR_C values above zero. The highest observed LR_C for an adventitious match was 93,665. The counts of the log(LR_C) values for all matches are provided as a summary in Fig. 1A.

The three person artificial profile both without and with dropout matched the three known contributors with the highest LR_C as expected. The LR_C values for all adventitious matches ($N = 16$ for no dropout and $N = 111{,}638$ for dropout) above zero are summarised Fig. 1B and C for no dropout and with dropout, respectively. The highest observed LR_C for an adventitious match was 5189 for the three person profile without dropout and 15,141 for the three person profile with dropout.

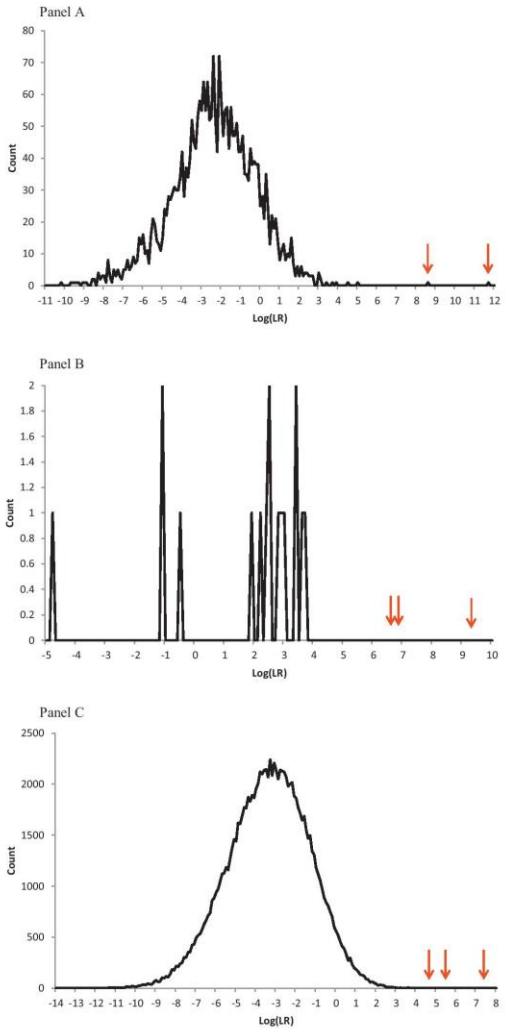


Fig. 1. Summary of counts of log(LR_C) values for all adventitious matches for the two person profile with dropout (A), three person profile without dropout (B) and three person profile with dropout (C). The log(LR_C) of the known contributors is indicated by arrows.

3.2. Experiment 2

A representative epg from one of the eight artificial mixed DNA profiles (Profile 1) is given within the supplementary material. The LR_C values for all adventitious matches are summarised in Table 1. The incorrect assignment of a fourth contributor to the interpretation generates many more possible genotype combinations and results in a large increase in the number of low grade adventitious links ($LR_C < 1000$). There was no trend observed which could be attributed to the number of contributors for adventitious links with $LR_C > 1000$. The highest observed LR_C for an adventitious match with either $N = 3$ (correct) or $N = 4$ (incorrect) was 114,000 (Profile 3, interpreted incorrectly as a four person mixture).

The LR_C for each of the known contributors considered individually as potential contributors to the artificial mixtures under H_1 is given in Table 2. Interpretation of the profile incorrectly assuming four contributors had little impact on the LR_C where the known individual was a major contributor. For minor contributors however, interpretation of the profile assuming four contributors had the effect of reducing the LR_C, in some cases up to three orders of magnitude, when compared to the LR_C calculated using the true number of contributors. As expected, comparison to the interpretation assuming three contributors resulted in the highest *LR*s for all profiles.

3.3. Experiment 3

A representative epg from one of the eight artificial mixed DNA profiles (Profile 1) reduced in scale by 90% is given in Fig. 2, supplementary material. The LR_C values for all adventitious matches are summarised as counts in Table 3. As in experiment 2, the assumption of more contributors to the profile results in many more possible genotype combinations. Fewer adventitious matches were observed when an assumption of two contributors was made. The adventitious match with the highest LR_C (730,000) occurred when Profile 8 was interpreted as a three person mixture. More adventitious matches with high LR_C values (in the order of 10^5) were obtained for the extreme low level profiles compared to experiment 2.

The LR_C for each of the known contributors to the 90% scaled mixtures is given in Table 4. The scaling of the profiles downwards by 90% resulted in the complete dropout of the trace contributor to each profile (highlighted in Table 4). As expected, the *LR* values for the known contributors are lower than the original comparison (Table 2) because of the increased uncertainty in the profile interpretation. As in experiment 2, comparison to the known contributors resulted in the highest LR_C values for all profiles.

3.4. Experiment 4

The artificially constructed four person mixture interpreted incorrectly as a three contributor profile linked to the three known 'major' contributors with *LR*s of 1.5×10^{11}, 1.1×10^5 and 4.9×10^{13}. These *LR*s are in within one order of magnitude of the original profile interpretation results in Table 2 indicating that the introduction of a trace contributor to a profile has little effect on the interpretation of the major profiles. The LR_C for the highest adventitious match was 4260.

The profile, when interpreted correctly as a four contributor profile, linked to the three known 'major' contributors with *LR*s of 5.5×10^{11}, 6.5×10^3 and 1.3×10^{14}. The additional fourth contributor matched to the corresponding database profile with LR_C of 11.6.

Table 1
Count of adventitious links per profile for experiment 2, a true three person mixture interpreted assuming either three or four contributors.

Profile	1		2		3		4		5		6		7		8		Total counts	
Assumed no. contributors	3	4	3	4	3	4	3	4	3	4	3	4	3	4	3	4	3	4
Ranges of LR_C $1–10^1$	3076	31,464	1209	4057	45	16,956	1	23,582	22	24,433	330	26,303	254	24,781	203	29,685	5140	181,261
$10^1–10^2$	960	3036	497	164	32	2717	31	3319	105	3678	287	2850	152	2777	826	3845	2890	22,386
$10^2–10^3$	168	125	123	2	10	196	43	137	120	102	85	123	36	191	301	139	886	1015
$10^3–10^4$	17	2	22	0	3	18	24	1	31	0	15	2	15	22	15	4	142	49
$10^4–10^5$	2	1	2	0	1	1	0	0	0	0	0	0	3	4	3	0	11	6
$10^5–10^6$	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
Total as % of database size	2.9%	23.8%	1.3%	2.9%	0.1%	13.7%	0.1%	18.6%	0.2%	19.4%	0.5%	20.1%	0.3%	19.1%	0.9%	23.1%	9069	204,718

Table 2
LR_C for known contributors to each of the artificial mixtures, experiment 2.

Sample	1		2		3		4		5		6		7		8		
Assumed no. contributors	3	4	3	4	3	4	3	4	3	4	3	4	3	4	3	4	
LRs for the three known contributors	1.4×10^{12}	2.0×10^{12}	3.2×10^{6}	9.7×10^{3}	4.8×10^{12}	3.0×10^{12}	1.7×10^{17}	1.5×10^{17}	1.3×10^{18}	1.3×10^{18}	1.4×10^{16}	7.7×10^{16}	1.4×10^{9}	5.8×10^{7}	1.4×10^{17}	1.6×10^{17}	
	4.1×10^{5}	5.8×10^{3}	2.3×10^{16}	2.3×10^{16}	1.7×10^{8}	1.2×10^{5}	2.3×10^{15}	2.4×10^{15}	1.7×10^{16}	1.8×10^{16}	7.0×10^{13}	6.4×10^{14}	2.3×10^{16}	2.3×10^{16}	1.2×10^{15}	1.3×10^{15}	
	6.4×10^{13}	8.2×10^{13}	8.0×10^{15}	5.9×10^{15}	1.1×10^{14}	1.0×10^{14}	3.3×10^{5}	4.6×10^{3}	1.0×10^{5}	1.6×10^{3}	2.1×10^{6}	3.5×10^{2}	1.9×10^{16}	5.2×10^{16}	1.4×10^{5}	3.7×10^{3}	

Table 3
Count of adventitious links per profile for the extreme low level profiles for experiment 3.

Profile		1		2		3		4		5		6		7		8		Total counts	
Assumed no. contributors		2	3	2	3	2	3	2	3	2	3	2	3	2	3	2	3	2	3
Ranges of LR_C	>1	7	1039	28	551	8	1181	7	755	3	864	1	930	8	596	1	749	63	6665
	10^1–10^2	25	143	20	216	9	40	6	111	3	64	1	0	0	328	1	122	65	1024
	10^2–10^3	6	18	4	34	2	9	5	15	2	13	3	5	2	54	3	9	27	157
	10^3–10^4	1	5	0	1	2	2	1	5	0	6	0	1	3	7	1	5	8	32
	10^4–10^5	1	1	0	0	0	1	0	0	0	0	0	1	0	0	0	1	1	4
	10^5–10^6	1	1	0	0	0	0	0	1	0	1	1	1	0	0	1	1	3	5
Total as % of database size		2.9%	84.5%	3.6%	56.1%	1.5%	86.3%	1.3%	62.1%	0.6%	66.3%	0.4%	65.6%	0.9%	68.9%	0.5%	62.1%	167	7887

Table 4
LR_C for known contributors to each of the extreme low level artificial mixtures, experiment 3 (trace contributor highlighted).

Sample	1		2		3		4		5		6		7		8	
Assumed no. contributors	2	3	2	3	2	3	2	3	2	3	2	3	2	3	2	3
LRs for the three known contributors	1.6×10^7	3.8×10^6	1.1	1.4	1.6×10^9	4.5×10^8	7.8×10^8	2.2×10^8	1.5×10^{13}	5.5×10^{10}	1.6×10^{11}	1.8×10^{11}	0	2.8	7.6×10^{10}	1.3×10^{11}
	0	4.6	1.6×10^{15}	1.5×10^{15}	0	2.3	9.5×10^{11}	2.1×10^{12}	4.1×10^{14}	3.4×10^{13}	1.1×10^8	4.1×10^7	1.8×10^{16}	1.7×10^{16}	9.7×10^7	4.0×10^7
	3.8×10^{11}	8.7×10^{10}	1.7×10^8	2.3×10^5	7.4×10^{11}	6.4×10^{11}	0	3.1	0	2.0	0	1.2	3.5×10^8	9.7×10^7	0	4.1

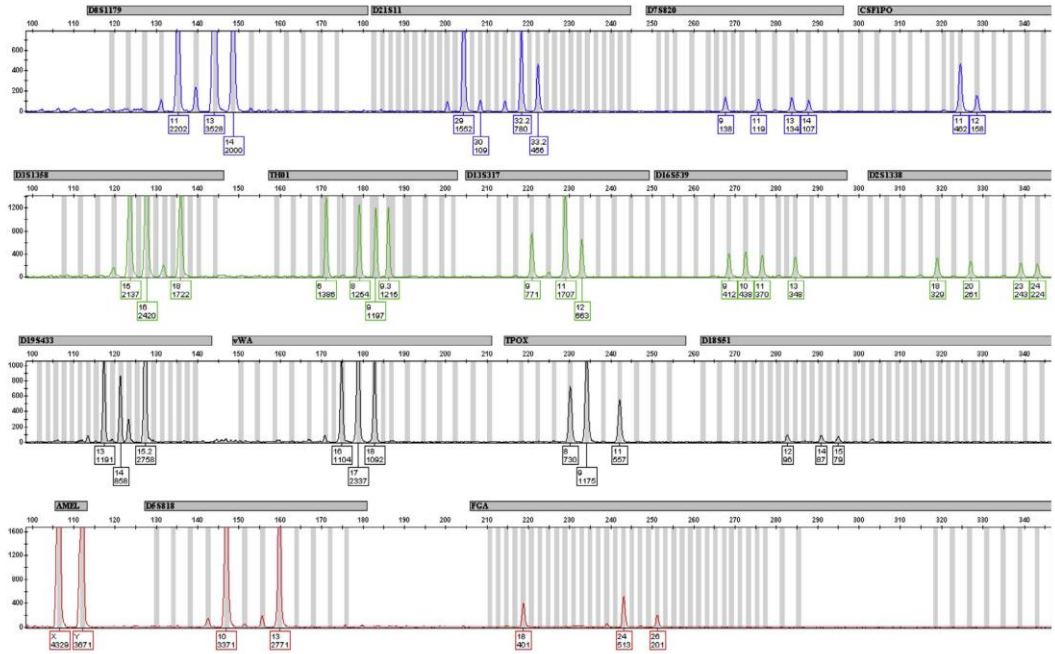


Fig. 2. Epg of mixed DNA profile from a semen stain from Case 1.

Case 1. A mixed DNA profile was obtained from a semen stain on a carpet at the scene of an alleged sexual assault involving two male offenders. DNA from most likely two contributors was detected, present in approximately equal proportions. The epg is shown in Fig. 2. The profile was interpreted assuming two contributors and searched against the DPD using a *LR* threshold of one million (10^6). The threshold was determined by rounding upwards the *LR* from the highest observed adventitious match in experiment 3. Each of the individuals on the database was considered as a potential contributor in turn under the following two hypotheses:

H_1. Database individual under consideration and one unknown contributor.

H_2. Two unknown contributors.

The crime profile was linked to two individuals. The profiles of the two individuals are in Table 5. Direct comparison of the individual profiles to the crime profile reveals a potential non-concordance with Contributor 2 at D18S51. On close inspection of the epg in Fig. 2 a peak in the 17 allele bin is visible below the analytical threshold. Despite this non-concordance and the large imbalance at D21S11, we note that these two contributors fully explain the complete profile.

Case 2. A low level mixed DNA profile was obtained from cellular material recovered from a shoe that was located in car at the scene

Table 5
Profiles of the two matching individuals and LR_C for Case 1.

Locus	Contributor 1	Contributor 2
D8	11, 14	13, 13
D21	30, 32.2	29, 33.2
D7	13, 14	9, 11
CSF	11, 12	11, 11
D3	15, 18	16, 16
TH01	8, 9.3	6, 9
D13	11, 11	9, 12
D16	10, 13	9, 11
D2	18, 24	20, 23
D19	15.2, 15.2	13, 14
vWA	17, 18	16, 17
TPOX	9, 11	8, 9
D18	14, 15	12, 17
D5	10, 13	10, 13
FGA	24, 26	18, 24
LR_C	4.9×10^{13}	9.3×10^{9}

Table 6
Profile of the matching individual and LR_C for Case 2.

Locus	Contributor 1
D8	14, 14
D21	30, 30
D7	–
CSF	–
D3	15, 17
TH01	7, 9.3
D13	–
D16	9, 12
D2	16, 23
D19	13, 14
vWA	14, 19
TPOX	–
D18	15, 17
D5	–
FGA	24, 26
LR_C	9.1×10^{8}

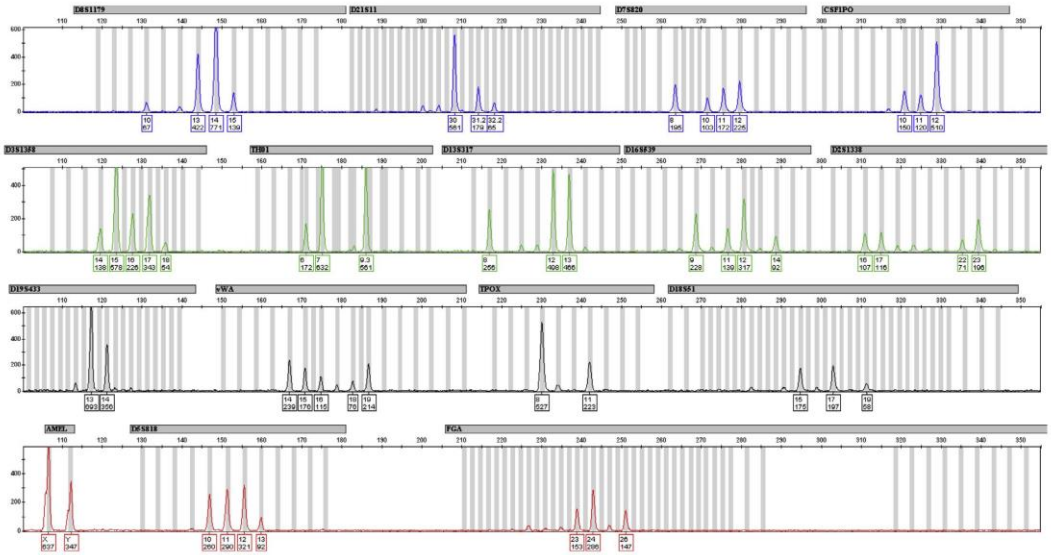


Fig. 3. Epg of the mixed DNA profile from a shoe insole from Case 2.

of an aggravated burglary. The epg of one of the replicate amplifications is shown in Fig. 3. The profile was interpreted assuming three contributors based on the minimum peak count, and supported by sub-threshold peak information. Each of the individuals on the database was considered as a potential contributor in turn under the following two hypotheses:

H_1. Database individual under consideration and two unknown contributor, and

H_2. Three unknown contributors.

The profile linked to one individual profiled using the SGMPlus multiplex on the DPD. The DNA profile of that individual and the corresponding LR_C is in Table 6.

4. Conclusion

Direct searching of unresolved mixtures against databases of known individuals has been shown to be feasible as an investigative technique with the use of a suitable *LR* threshold to filter out low grade adventitious links. For this dataset, an appropriate LR_C threshold of approximately 1 million would ensure the risk of reporting an adventitious match is mitigated when interpreting extreme low level profiles (the majority of peaks less than 400 rfu and all peaks below 800 rfu). Complex DNA profiles with different contributors in the same proportions resulted in the highest LR_C values when the known contributors were considered individually as potential contributors under H_1, even when dropout was a consideration. The choice of a threshold is undertaken as part of a risk assessment. Setting the threshold too low risks increasing the chance of obtaining an adventitious match whereas setting the threshold too high risks missing true, legitimate matches. Table 7 shows the rate of adventitious matches (false positives) and incorrect non-matches (false negatives) that arise from using different *LR* cut off values using data from Tables 1 and 2.

Regardless of where the search threshold is set there will always be the possibility of false positive and false negative results. There is a limited capability to identify a true contributor if they are a trace contributor to a complex mixture without also flagging a large number of false positive links. This is also true if there is substantial dropout of an individual's alleles or if a minor contributor's alleles are masked by a major contributor's alleles

Table 7
Numbers of false negative and false positive results obtained in experiment 2 using different *LR* cut off values.

LR cut off	Considered a 3 person mix		Considered a 4 person mix	
	Number of false positives	Number of false negatives	Number of false positives	Number of false negatives
10^6	0	4	0	7
10^5	0	0	1	6
10^4	11	0	7	6
10^3	153	0	56	1
10^2	1039	0	1071	0
10^1	3929	0	23,457	0
1	9069	0	204,718	0

within a mixed DNA profile. This information can be used to help form guidelines in order to limit the numbers of mixed DNA profiles searched against a database to those that have the greatest potential to provide strong investigative leads.

The assumption of additional contributors to a profile beyond those suggested by allele count alone tended to lower the LR_C for the true minor and major contributors and increase the number of low grade adventitious links, where $1 < LR_C < 1000$. A match against the database is unlikely for a trace contributor that has very few alleles either present above the analytical threshold and present in non-masked allele positions. This is the expected outcome.

The assumption of additional contributors also resulted in significantly increased computational effort.

The multiplex used to determine the genotype for the known database profile did not appear to have an effect on whether an adventitious link was made. This was evident from the make-up of the high risk profile database where the profile multiplexes were in the same appropriate proportions of the original database.

Two case examples are described where profiles that were considered previously unsuitable for database comparison were interpreted and searched against the NZ DPD with a LR_C threshold of 1 million. Both cases resulted in links to individuals with high LR_C values. It is worth cautioning the reader that, as with any links resulting from a database search, their primary purpose in investigative only and further scrutiny is warranted.

This work reinforces the power of DNA databases as investigative tools and demonstrates the ability to directly search mixed DNA profiles using a *LR* framework without the need to identify a single contributor profile. Even using fully continuous methods of interpretation an individual's profile may not be able to be reasonably inferred from a complex DNA mixture. This would make the profile unsuitable for entry to a database using traditional database search methods. The searching method proposed in this paper allows real time searching of complex mixed DNA profiles. Using an *LR* strategy is a more powerful method than counting matching alleles, for example, and allows phenomena such as drop in and dropout to be taken into account. The functionality is available on the NZ DPD where average search times are 10 min against a database of over 145,000 profiles.

Acknowledgements

This work was supported in part by grant 2011-DN-BX-K541 from the US National Institute of Justice. Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the U.S. Department of Justice. The authors thank Catherine McGovern and Johanna Veth (ESR) and two anonymous reviewers whose helpful comments greatly improved this paper.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.fsigen.2013.12.001.

References

[1] S.J. Walsh, J.S. Buckleton, DNA intelligence databases, in: J.S. Buckleton, M.C. Triggs, S.J. Walsh (Eds.), Forensic DNA Evidence Interpretation, CRC Press, Boca Raton, 2005.
[2] K. Devlin, New DNA technique could help reopen thousands of 'cold cases', in: The Telegraph, 23 March, 2009.
[3] L. Mueller, Can simple population genetic models reconcile partial match frequencies observed in large forensic databases? J. Genet. 87 (2) (2008).
[4] J.M. Butler, DNA databases: uses and issues, in: J.M. Butler (Ed.), Advanced Topics in Forensic DNA Typing: Methodology, Elsevier, 2012.
[5] P. Gill, C.H. Brenner, J.S. Buckleton, A. Carracedo, M. Krawczak, W.R. Mayr, et al., DNA commission of the International Society of Forensic Genetics: recommendations on the interpretation of mixtures, Forensic Sci. Int. 160 (2–3) (2006) 90–101.
[6] F.R. Bieber, C.H. Brenner, D. Lazer, Finding criminals through DNA of their relatives, Science 312 (2006) 1315–1316.
[7] D.J. Balding, M. Krawczak, J.S. Buckleton, J.M. Curran, Decision-making in familial database searching: KI alone or not alone? Forensic Sci. Int.: Genet. 7 (1) (2013) 52–54.
[8] B. Caddy, G. Taylor, A. Linacre, A review of the science of low template DNA analysis, 2008 Available from: http://police.homeoffice.gov.uk/news-and-publications/publication/operational-policing/Review_of_Low_Template_DNA_1.pdf?view=Binary (cited 14th April 2008).
[9] The Forensic Science Regulator, Response to Professor Brian Caddy's Review of the Science of Low Template DNA Analysis, 2008 Available from: http://police.homeoffice.gov.uk/news-and-publications/publication/operational-policing/Review_of_Low_Template_DNA_1.pdf?view=Binary (cited 12 May 2008).
[10] J.-A. Bright, D. Taylor, J.M. Curran, J.S. Buckleton, Developing allelic and stutter peak height models for a continuous method of DNA interpretation, Forensic Sci. Int.: Genet. 7 (2) (2013) 296–304.
[11] H. Haned, L. Pène, J.R. Lobry, A.B. Dufour, D. Pontier, Estimating the number of contributors to forensic DNA mixtures: does maximum likelihood perform better than maximum allele count? J. Forensic Sci. 56 (1) (2011) 23–28.
[12] A. Biedermann, S. Bozza, K. Konis, F. Taroni, Inference about the number of contributors to a DNA mixture: comparative analyses of a Bayesian network approach and the maximum allele count method, Forensic Sci. Int.: Genet. 6 (6) (2012) 689–696.
[13] B. Budowle, A.J. Onorato, T.F. Callaghan, A.D. Manna, A.M. Gross, R.A. Guerrieri, et al., Mixture interpretation: defining the relevant features for guidelines for the assessment of mixed DNA profiles in forensic casework, J. Forensic Sci. 54 (4) (2009) 810–821.
[14] J.-A. Bright, P. Gill, J. Buckleton, Composite profiles in DNA analysis, Forensic Sci. Int.: Genet. 6 (3) (2012) 317–321.
[15] H. Kelly, J.-A. Bright, J. Curran, J. Buckleton, The interpretation of low level DNA mixtures, Forensic Sci. Int.: Genet. 6 (2) (2012) 191–197.
[16] T.M. Clayton, J.S. Buckleton, Mixtures, in: J.S. Buckleton, C.M. Triggs, S.J. Walsh (Eds.), Forensic DNA Evidence Interpretation, CRC Press, Boca Raton, Florida, 2004 pp. 217–274.
[17] H. Kelly, J.A. Bright, J.S. Buckleton, J.M. Curran, A comparison of statistical models for the analysis of complex forensic DNA profiles, Sci. Justice (2013) http://www.scopus.com/inward/record.url?eid=2-s2.0-84881142115&partnerID=40&md5=48ba7495cff95503c481dde967ae91d3.
[18] D. Taylor, J.A. Bright, J.S. Buckleton, The interpretation of single source and mixed DNA profiles, Forensic Sci. Int.: Genet. 7 (2013) 516–528.
[19] J.A. Bright, D. Taylor, J.M. Curran, J.S. Buckleton, Degradation of forensic DNA profiles, Aust. J. Forensic Sci. (2013). http://dx.doi.org/10.1080/00450618.2013.772235.
[20] M.W. Perlin, M.M. Legler, C.E. Spencer, J.L. Smith, W.P. Allan, J.L. Belrose, et al., Validating TrueAllele® DNA mixture interpretation, J. Forensic Sci. 56 (6) (2011) 1430–1447.
[21] R. Puch-Solis, L. Rodgers, A. Mazumder, S. Pope, I. Evett, J. Curran, et al., Evaluating forensic DNA profiles using peak heights, allowing for multiple donors, allelic dropout and stutters, Forensic Sci. Int.: Genet. 7 (5) (2013) 555–563.
[22] J.S. Buckleton, J.M. Curran, P. Gill, Towards understanding the effect of uncertainty in the number of contributors to DNA stains, Forensic Sci. Int.: Genet. 1 (1) (2007) 20–28.
[23] S.A. Harbison, J.F. Hamilton, S.J. Walsh, The New Zealand DNA databank: its development and significance as a crime solving tool, Sci. Justice 41 (2001) 33–37.
[24] J.-A. Bright, J.S. Buckleton, C.E. McGovern, Allele frequencies for the four major sub-populations of New Zealand for the 15 Identifiler loci, Forensic Sci. Int.: Genet. 4 (2) (2010) e65–e66.
[25] Applied Biosystems, AmpFlSTR Identifiler Amplification Kit User Guide, Applied Biosystems, Foster City, CA, 2009.
[26] D. Taylor, J.-A. Bright, J.S. Buckleton, The interpretation of single source and mixed DNA profiles, Forensic Sci. Int.: Genet. 7 (5) (2013) 516–528.

Manuscript: Considering relatives when assessing the evidential strength of mixed DNA profiles. D Taylor, JA Bright, J Buckleton. (2014) Forensic Science International: Genetics 13, 259-263 – *Cited 4 times*

Statement of novelty: This work derives the mathematics that extends the LR theory to complex mixtures when considering relatives of a person of interest as a potential contributor. This extended the existing theory (which was mostly focussed on single source profiles) and provided mathematical derivations to explain the extension. In this work we also demonstrated the idea of carrying out familial searches against complex mixtures, using an extension of the same theory.

My contribution: Main theorist and writer of the work.

Research Design / Data Collection / Writing and Editing = 60% / NA / 70%

Additional comments:

Forensic Science International: Genetics 13 (2014) 259–263

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

Original Research

Considering relatives when assessing the evidential strength of mixed DNA profiles

Duncan Taylor [a,b,*], Jo-Anne Bright [c], John Buckleton [c]

[a] Forensic Science South Australia, 21 Divett Place, Adelaide, SA 5000, Australia
[b] School of Biological Sciences, Flinders University, GPO Box 2100, Adelaide, SA 5001, Australia
[c] ESR, Private Bag 92021, Auckland 1142, New Zealand

ARTICLE INFO

Article history:
Received 15 April 2014
Received in revised form 28 July 2014
Accepted 31 August 2014

Keywords:
DNA profile interpretation
Likelihood ratio
Mixtures
Relatives
Continuous model
STRmix

ABSTRACT

Sophisticated methods of DNA profile interpretation have enabled scientists to calculate weights for genotype sets proposed to explain some observed data. Using standard formulae these weights can be incorporated into an *LR* calculation that considers two competing propositions. We demonstrate here how consideration of relatedness to the person of interest can be incorporated into a *LR* calculation and how the same calculation can be used for familial searches of complex mixtures. We provide a general formula that can be used in semi or fully automated methods of calculation and demonstrate their use by working through an example.

1. Introduction

The introduction of new forensic DNA profiling kits has resulted in improved discrimination power from an increased number of loci and an increase in sensitivity due to improvements in kit chemistry [1,2]. Methods of evaluating DNA profiles have also benefitted from recent improvements, allowing the interpretation of many more profiles and the generation of an evidential weight, in the form of a likelihood ratio (*LR*), for the comparison of evidence to reference profiles [3–6]. The *LR* considers the probability of the evidence under two propositions aligned to the prosecution and defence cases, respectively. Often the proposition considered for the defence is that the DNA profile has originated from other unrelated individuals.

Whilst such an *LR* addresses the question most often asked by the court, the issue of relatives does arise and is plausibly not given the attention it deserves [7]. Specifically, it may be of interest whether a relative of the nominated person of interest (POI) was a contributor to the DNA sample rather than the POI themselves [8]. There are many situations where a reference sample of relatives may not be available and so the question of their contribution cannot be answered biologically, by direct comparison to the crime profiles.

We investigate here the consideration of relatives as an alternative source of DNA in mixed DNA profiles. In particular we take into account the effect of profile complexity, contributor order, coancestry (Fst), and alleles that are identical by decent (IBD) on the calculation.

We present a general formula that can be used to calculate an *LR* considering relatives and use this general formula on a worked example.

2. General mathematics

Consider a profile originating from n individuals. A nominated POI contains all the necessary alleles so that they explain one component of the mixture. The prosecution asserts that the DNA profile originated from the POI and $n-1$ other, unrelated individuals. The defence assert that the POI is not a contributor and instead the DNA sample has originated from n other individuals, one of whom is related to the POI. The propositions that are being considered in the calculation of the *LR* are:

H_p: the DNA sample originated from the POI and $n-1$ other, unrelated individuals.
H_d: the DNA sample originated from a relative of the POI and $n-1$ other, unrelated individuals.

* Corresponding author. Tel.: +61 8 8226 7700; fax: +61 8 8226 7777.
E-mail address: Duncan.Taylor@sa.gov.au (D. Taylor).

http://dx.doi.org/10.1016/j.fsigen.2014.08.015
1872-4973/

In this work we make the simplification that we consider *LR*s where there is only one person of interest. This simplification is not required mathematically, however the consideration of alternate propositions that involve combinations of relatives becomes complex. Considering problems of this type could prove challenging to the courts, as there is no clear indication as to what the alternative proposition is.

Starting generally, we consider some observed peak data (**O**):

$$\mathrm{LR} = \frac{p(\boldsymbol{O}|H_p)}{p(\boldsymbol{O}|H_d)} \tag{1}$$

The DNA profile is analysed at each locus and a list of j genotype sets ($\mathbf{S_j}$) are obtained that could explain the observed peaks (**O**), each with an associated weight (w_j) which relates to the probability density of **O** if $\mathbf{S_j}$ was the genotype set that gave rise to it, i.e. $p(O|S_j) \propto w_j$, so that Eq. (1) becomes:

$$\mathrm{LR} = \frac{\sum_j w_j \Pr(\boldsymbol{S_j}|H_p)}{\sum_{j'} w_{j'} \Pr(\boldsymbol{S_{j'}}|H_d)} \tag{2}$$

where j is the number of genotype sets being considered under H_p and j' is the number of genotype sets being considered under H_d. The *LR* is therefore the weighted sum of genotype sets applicable under each proposition.

We present in Eq. (3) a general formula for the *LR* that considers a relative as an alternative source of DNA to the POI in a mixed DNA sample. Eq. (3) does not consider sub-population effects. A complete derivation of Eq. (3) in generality is given in appendix 1, which includes a sub-population effect.

$$\mathrm{LR} = \frac{\sum_i \sum_j w_j \Pr(\boldsymbol{S_{U_j}}|\mathbb{C}_i, H_p)}{\sum_i \sum_{j'} w_{j'} \Pr(G_{R_{j'}}|G_p, \mathbb{C}_i, H_d) \Pr(\boldsymbol{S_{U_{j'}}}|\mathbb{C}_i, H_d)} \tag{3}$$

Eq. (3) introduces some new terms. G signifies the genotype of a single contributor, with the right subscript indicating the person referred to. Using this nomenclature we designate G_p as the genotype of the POI and $G_{R_{j'}}$ as the genotype of the individual related to the POI. As the genotype being considered for the relative changes depending on the genotype set being considered it also has a sub-sub script j'. In Eq. (3) we also introduce contributor order ($\mathbb{C}$). Contributor order takes into account the fact that we do not nominate a specific component of the mixture to compare to the POI in the propositions, i.e. if a POI was being compared to a two person mixture with a major and minor component then they could be compared to either component of the mixture because *a priori* there is no reason to choose one over the other. Consider the example where the POI is the source of the major component. If an *LR* was then presented that ignored the fact that the POI was excluded as the minor contributor then this is not providing a result for the comparison of the POI to the profile, but rather only a single component of it. When considering relatives as alternative sources of DNA then contributor order is equally important, and arguably more intuitive. If considering a relative of a POI to the same two person mixture as described above there is no need to restrict the consideration of the relative to the same component of the mixture that matched the POI.

The probabilities of genotypes of unknown contributors, S_U, are calculated in the standard manner as published by Balding and Nichols in 1994 [10] using the sampling formula. The probability of $G_{R_{j'}}$, with genotype $G_{R_{j'}} = [a_1, a_2]$, given G_P, with genotype $G_P = [a_3, a_4]$, can be calculated using a single standard formula, making use of indicator terms, α (an explanation follows shortly):

$$\Pr(G_{R_{j'}}|G_p, \mathbb{C}_i, H_d) = \alpha_2 \Pr(Z_2) + \frac{\Pr(Z_1)}{2}\left(\alpha_{1B}\, p_{a_1} + \alpha_{1A}\, p_{a_2}\right) + \alpha_0 \Pr(Z_0)\, p_{a_1}\, p_{a_2} \tag{4}$$

Table 1
Probability that two individuals with a given relationship share 0, 1 or 2 IBD alleles.

Relationship	$\Pr(Z_0)$	$\Pr(Z_1)$	$\Pr(Z_2)$
Unrelated	1	0	0
Full-sibling	¼	½	¼
Half-sibling	½	½	0
Parent/Child	0	1	0
First cousin	¾	¼	0

Eq. (4) includes probabilities that either both, one or none of the alleles between the $G_{R_{j'}}$ and G_p are identical by decent (IBD) with $\Pr(Z_2)$, $\Pr(Z_1)$, and $\Pr(Z_0)$ respectively. Values of $\Pr(Z_x)$ are available from many sources for commonly considered relationships and we provide a reproduction of that information in Table 1 [10,11].

In Eq. (4) we also use indicator terms, α, which allow the use of a single formula. These terms take into account the genotypes of G_P and $G_{R_{j'}}$ and their ability to possess 0, 1 or 2 IBD alleles. Table 2 shows the values that the indicator terms will take for different combinations of G_P and $G_{R_{j'}}$.

2.1. Example

We now introduce a visual representation of one locus of a two person mixed DNA profile in Fig. 1, to aid explanations.

The genotype sets and weights associated with this DNA profile are given in Table 2. As we are taking contributor order into account we use a left superscript to signify contributor position.

The person if interest has genotype [13,14] for this locus. We initially consider the propositions:

H_p: the DNA sample originated from the POI and $n-1$ other, unrelated individuals.
H_{d1}: the DNA sample originated from n individuals, unrelated to the POI.

We first calculate $p(O|H_p)$ by calculating $\sum_i \sum_j w_j \Pr(S_{U_j}|\mathbb{C}_i, H_p)$ from Eq. (3). For a two person mixture there are two contributor orders, meaning:

$$p(O|H_p) = \sum_j w_j \Pr(S_{U_j}|\mathbb{C}_1, H_p) + \sum_j w_j \Pr(S_{U_j}|\mathbb{C}_2, H_p) = 0.55 \times 2\, p_{16}\, p_{17} + 0.05 \times 2\, p_{16}\, p_{17}$$

Table 2
Value for indicator terms in Eq. (4) given genotypes $G_{R_{j'}} = [a_1, a_2]$ and $G_P = [a_3, a_4]$.

Indicator term	Value	Condition
α_2	1	$a_1 = a_3$ and $a_2 = a_4$
	0	$a_1 \neq a_3$ or $a_2 \neq a_4$
α_{1A}	1	$a_1 = a_3$ or $a_1 = a_4$
	0	$a_1 \neq a_3$ and $a_1 \neq a_4$
α_{1B}	1	$a_2 = a_3$ or $a_2 = a_4$
	0	$a_2 \neq a_3$ and $a_2 \neq a_4$
α_0	1	$a_1 = a_2$
	2	$a_1 \neq a_2$

Weights for genotype sets under for Fig. 1.

Set (j)	Genotype set ($\mathbf{S_j}$)		Weight (w_j)
	${}^1G_j$	${}^2G_j$	
1	[13,14]	[16,17]	0.55
2	[13,16]	[14,17]	0.10
3	[13,17]	[14,16]	0.10
4	[14,16]	[13,17]	0.10
5	[14,17]	[13,16]	0.10
6	[16,17]	[13,14]	0.05

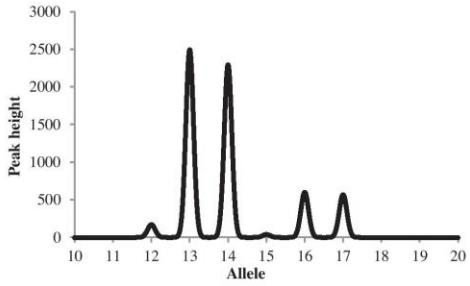


Fig. 1. Example electropherogram showing a single locus of a two person mixture.

Where the first term considers the POI in contributor position 1 and the second term in contributor position 2. Simplification provides:

$$p(\boldsymbol{O}|H_p) = 1.2 \times p_{16}\,p_{17}$$

Using standard formulation:

$$p(\boldsymbol{O}|H_{d1}) = \sum_i \sum_j w_j \Pr(\boldsymbol{S_{U_j}}|\mathbb{C}_i, H_{d1}) = \sum_j w_j \Pr(\boldsymbol{S_{U_j}}|\mathbb{C}_1, H_{d1}) + \sum_j w_j \Pr(\boldsymbol{S_{U_j}}|\mathbb{C}_2, H_{d1})$$

Expanding Table 2 to incorporate calculation elements provides the results seen in Table 3.

Therefore:

$$\sum_j w_j \Pr(\boldsymbol{S_{U_j}}|\mathbb{C}_1, H_{d1}) + \sum_j w_j \Pr(\boldsymbol{S_{U_j}}|\mathbb{C}_2, H_{d1}) = 8\,p_{13}\,p_{14}\,p_{16}\,p_{17}$$

So that the *LR* is:

$$\mathrm{LR}_1 = \frac{p(\boldsymbol{O}|H_p)}{p(\boldsymbol{O}|H_{d1})} = \frac{1.2 \times p_{16}\,p_{17}}{8\,p_{13}\,p_{14}\,p_{16}\,p_{17}} = \frac{3}{20\,p_{13}\,p_{14}}$$

Now we consider the proposition:

H_{d2}: the DNA sample originated from a sibling of the POI and $n-1$ other, unrelated individuals

From Eq. (3) we wish to calculate $\sum_i \sum_{j'} w_{j'} \Pr(G_{R_{j'}}|G_p, \mathbb{C}_i, H_{d2})\Pr(S_{U_{j'}}|\mathbb{C}_i, H_{d2})$ where we have removed independent terms as a result of ignoring substructure. We also wish to use Eq. (4) to calculate the *LR*. We expand Table 3 to include elements of Eqs. (3) and (4) and display them in Table 4. We wish to calculate:

$$\sum_i \sum_{j'} w_{j'} \Pr(G_{R_{j'}}|G_p, \mathbb{C}_i, H_d)\Pr(S_{U_{j'}}|\mathbb{C}_i, H_d)$$

These terms can be obtained by summing the multiplied element from Table 4 so that:

$$\sum_i \sum_{j'} w_{j'} \Pr(G_{R_{j'}}|G_p, \mathbb{C}_i, H_d)\Pr(G_{U_{j'}}|\mathbb{C}_i, H_d) = [2\,p_{16}\,p_{17}]\left\{\begin{array}{l} 0.6\left[\frac{1}{4}+\frac{1}{4}(p_{13}+p_{14})+\frac{1}{4}2\,p_{13}\,p_{14}\right]+ \\ 0.2\left[\frac{1}{4}p_{14}+\frac{1}{4}2\,p_{13}\,p_{14}\right]+ \\ 0.2\left[\frac{1}{4}p_{14}+\frac{1}{4}2\,p_{13}\,p_{14}\right]+ \\ 0.2\left[\frac{1}{4}p_{13}+\frac{1}{4}2\,p_{14}\,p_{13}\right]+ \\ 0.2\left[\frac{1}{4}p_{13}+\frac{1}{4}2\,p_{14}\,p_{13}\right]+ \\ 0.6\left[\frac{1}{4}2\,p_{13}\,p_{14}\right] \end{array}\right\}$$

and the *LR* can be calculated by:

$$\mathrm{LR}_2 = \frac{p(\boldsymbol{O}|H_p)}{p(\boldsymbol{O}|H_{d2})} = \frac{12}{3+5(p_{13}+p_{14})+20\,p_{13}\,p_{14}}$$

By substituting values $p_{13} = 0.003$ and $p_{14} = 0.1217$ (values for D3S1358 from [12]) we obtain $LR_1 \approx 410.8$ and $LR_2 \approx 3.3$, showing a considerable decrease from LR_1 to LR_2, when considering a close relative as the alternate source of DNA rather than an unrelated individual.

Table 3
Elements of calculation of $p(\boldsymbol{O}|H_{d1})$.

Set (j)	Genotype set ($\mathbf{S_j}$)		Weight (w_j)	$\Pr(\boldsymbol{S_{U_j}}\|\mathbb{C}_1, H_{d1})$	$\Pr(\boldsymbol{S_{U_j}}\|\mathbb{C}_2, H_{d1})$
	${}^1G_j$	${}^2G_j$			
1	[13,14]	[16,17]	0.55	$4\,p_{13}\,p_{14}\,p_{16}\,p_{17}$	$4\,p_{13}\,p_{14}\,p_{16}\,p_{17}$
2	[13,16]	[14,17]	0.10	$4\,p_{13}\,p_{14}\,p_{16}\,p_{17}$	$4\,p_{13}\,p_{14}\,p_{16}\,p_{17}$
3	[13,17]	[14,16]	0.10	$4\,p_{13}\,p_{14}\,p_{16}\,p_{17}$	$4\,p_{13}\,p_{14}\,p_{16}\,p_{17}$
4	[14,16]	[13,17]	0.10	$4\,p_{13}\,p_{14}\,p_{16}\,p_{17}$	$4\,p_{13}\,p_{14}\,p_{16}\,p_{17}$
5	[14,17]	[13,16]	0.10	$4\,p_{13}\,p_{14}\,p_{16}\,p_{17}$	$4\,p_{13}\,p_{14}\,p_{16}\,p_{17}$
6	[16,17]	[13,14]	0.05	$4\,p_{13}\,p_{14}\,p_{16}\,p_{17}$	$4\,p_{13}\,p_{14}\,p_{16}\,p_{17}$

Table 4
Elements of calculation of $p(O|H_{d2})$.

Set (j)	Genotype set ($\mathbf{S_j}$)		Weight (w_j)	$\mathbb{C}_1$ ($G_{R_{j'}} = {}^1G_j$)				$\mathbb{C}_2$ ($G_{R_{j'}} = {}^2G_j$)				$\Pr(Z_0)$	$\Pr(Z_{1A})$	$\Pr(Z_{1B})$	$\Pr(Z_2)$
	${}^1G_j$	${}^2G_j$		α_0	α_{1a}	α_{1b}	α_2	α_0	α_{1a}	α_{1b}	α_2				
1	[13,14]	[16,17]	0.55	2	1	1	1	2	0	0	0	¼	¼	¼	¼
2	[13,16]	[14,17]	0.10	2	1	0	0	2	0	1	0	¼	¼	¼	¼
3	[13,17]	[14,16]	0.10	2	1	0	0	2	0	1	0	¼	¼	¼	¼
4	[14,16]	[13,17]	0.10	2	0	1	0	2	1	0	0	¼	¼	¼	¼
5	[14,17]	[13,16]	0.10	2	0	1	0	2	1	0	0	¼	¼	¼	¼
6	[16,17]	[13,14]	0.05	2	0	0	0	2	1	1	1	¼	¼	¼	¼

Set (j)	Genotype set ($\mathbf{S_j}$)		Weight (w_j)	$\Pr(G_{R_{j'}}\|G_p, \mathbb{C}_1, H_{d2})$ i.e. $G_{R_{j'}} = {}^1G_j$	$\Pr(G_{U_j}\|\mathbb{C}_1, H_{d2})$ i.e. $G_{U_j} = {}^2G_j$	$\Pr(G_{U_j}\|\mathbb{C}_2, H_{d2})$ i.e. $G_{U_j} = {}^1G_j$	$\Pr(G_{R_{j'}}\|G_p, \mathbb{C}_2, H_{d2})$ i.e. $G_{R_{j'}} = {}^2G_j$
	${}^1G_j$	${}^2G_j$					
1	[13,14]	[16,17]	0.55	$\frac{1}{4}+\frac{1}{4}(p_{13}+p_{14})+\frac{1}{4}2\,p_{13}\,p_{14}$	$2\,p_{16}\,p_{17}$	$2\,p_{13}\,p_{14}$	$\frac{1}{4}2\,p_{16}\,p_{17}$
2	[13,16]	[14,17]	0.10	$\frac{1}{4}p_{16}+\frac{1}{4}2\,p_{13}\,p_{16}$	$2\,p_{14}\,p_{17}$	$2\,p_{13}\,p_{16}$	$\frac{1}{4}p_{17}+\frac{1}{4}2\,p_{14}\,p_{17}$
3	[13,17]	[14,16]	0.10	$\frac{1}{4}p_{17}+\frac{1}{4}2\,p_{13}\,p_{17}$	$2\,p_{14}\,p_{16}$	$2\,p_{13}\,p_{17}$	$\frac{1}{4}p_{16}+\frac{1}{4}2\,p_{14}\,p_{16}$
4	[14,16]	[13,17]	0.10	$\frac{1}{4}p_{16}+\frac{1}{4}2\,p_{14}\,p_{16}$	$2\,p_{13}\,p_{17}$	$2\,p_{14}\,p_{16}$	$\frac{1}{4}p_{17}+\frac{1}{4}2\,p_{13}\,p_{17}$
5	[14,17]	[13,16]	0.10	$\frac{1}{4}p_{17}+\frac{1}{4}2\,p_{14}\,p_{17}$	$2\,p_{13}\,p_{16}$	$2\,p_{14}\,p_{17}$	$\frac{1}{4}p_{16}+\frac{1}{4}2\,p_{13}\,p_{16}$
6	[16,17]	[13,14]	0.05	$\frac{1}{4}2\,p_{16}\,p_{17}$	$2\,p_{13}\,p_{14}$	$2\,p_{16}\,p_{17}$	$\frac{1}{4}+\frac{1}{4}(p_{13}+p_{14})+\frac{1}{4}2\,p_{13}\,p_{14}$

2.2. Familial searches

With a reasonably straightforward extension of the formula given in Section 2 a search of a complex mixture against a database can be carried out (as described in [13]), but seeking to weigh the proposition that a relative of the person in the database is a contributor to the mixture.

Consider an *LR* calculation where the defence proposition suggests an individual unrelated to the POI is the alternate source of DNA, so for an n person mixture we consider:

H_p: the DNA sample originated from the POI and $n-1$ other, unrelated individuals.
H_{d1}: the DNA sample originated from n individuals, unrelated to the POI.

We have described in this work the calculation of a set of propositions that instead considers a relative of the POI as an alternate source of DNA:

H_p: the DNA sample originated from the POI and $n-1$ other, unrelated individuals.
H_{d2}: the DNA sample originated from a relative of the POI and $n-1$ other, unrelated individuals.

By considering H_{d2} against H_{d1} the *LR* can be used in the same way that familial searches have been performed for single source profiles. Unlike standard familial searches, using the methodology described here allows such a search to be carried out against complex and unresolvable mixtures from which no single contributor's genotype can be unambiguously assigned.

3. Conclusion

The consideration of the possibility that a relative is a donor to a DNA stain has been shown previously to be of importance when considering single source stains [7]. Relatedness is likely to have a similar effect when considering an unresolvable mixture. The implementation of this consideration for mixtures does not require any new concepts. All that is required is a careful consideration of IBD states, coancestry, and contributor order. Noting that no new concepts are involved the combination is complex and tedious to apply. We have implemented this as a module in the STRmix™ software.

Acknowledgements

This work was supported in part by grant 2011-DN-BX-K541 from the US National Institute of Justice. Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the U.S. Department of Justice. We acknowledge the valuable comments of two anonymous referees that have greatly improved this manuscript.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.fsigen.2014.08.015.

Appendix C. Derivation of the *LR* formula

If references from individuals who are relevant to the case have been profiled we introduce them as $\mathbf{S_k}$ for known contributors (assumed to be present in the DNA profile by all parties) and $\mathbf{S_p}$ for the genotype sets of the POIs being postulated as contributors under one proposition but not all (typically in forensic contexts this will be a POI considered a contributor under H_p but not in H_d). By adding in genotype sets and weights to the *LR* Eq. (2) becomes:

$$\mathrm{LR}=\frac{\sum_j w_j \Pr(S_j|\boldsymbol{S_k},\boldsymbol{S_p},H_p)\Pr(\boldsymbol{S_k},\boldsymbol{S_p}|H_p)}{\sum_{j'} w_{j'}\Pr(\boldsymbol{S_{j'}}|\boldsymbol{S_k},\boldsymbol{S_p},H_d)\Pr(\boldsymbol{S_k},\boldsymbol{S_p}|H_d)}$$

The probabilities of genotypes for the known contributors $\mathbf{S_p}$ and $\mathbf{S_k}$ are equal under both proposition so that:

$$\mathrm{LR}=\frac{\sum_j w_j \Pr(\boldsymbol{S_j}|\boldsymbol{S_k},\boldsymbol{S_p},H_p)}{\sum_{j'} w_{j'}\Pr(\boldsymbol{S_{j'}}|\boldsymbol{S_k},\boldsymbol{S_p},H_d)}$$

Following Taylor et al. [9] we now introduce contributor order ($\mathbb{C}$). Also, in addition to the known contributors ($\mathbf{S_k}$) and the POIs ($\mathbf{S_p}$), the genotype sets ($\mathbf{S_j}$) can be further broken down into the genotypes of:

- Individuals related to the POI (S_{R_j}) and
- Unknown contributors ($\mathbf{S_U}$) that must be present in the mixture to explain **O** not accounted for by known contributors or POIs in the specified number of contributors.

This gives:
$\Pr(\boldsymbol{S_j}|\boldsymbol{S_k},\boldsymbol{S_p},H_p)=\Pr(\boldsymbol{S_{U_j}},\boldsymbol{S_{R_j}},\boldsymbol{S_p},\boldsymbol{S_k}|\boldsymbol{S_k},\boldsymbol{S_p},\mathbb{C}_i,H_p)$ and
$\Pr(\boldsymbol{S_j}|\boldsymbol{S_k},\boldsymbol{S_p},H_d)=\Pr(\boldsymbol{S_{U_j}},\boldsymbol{S_{R_j}},\boldsymbol{S_k}|\boldsymbol{S_k},\boldsymbol{S_p},\mathbb{C}_i,H_d)$
which leads to the *LR*:

$$\mathrm{LR}=\frac{\sum_i\sum_j w_j \Pr(\boldsymbol{S_{U_j}},\boldsymbol{S_{R_j}},\boldsymbol{S_p},\boldsymbol{S_k}|\boldsymbol{S_k},\boldsymbol{S_p},\mathbb{C}_i,H_p)\Pr(\mathbb{C}_i)}{\sum_i\sum_{j'} w_{j'}\Pr(\boldsymbol{S_{U_{j'}}},\boldsymbol{S_{R_{j'}}},\boldsymbol{S_k}|\boldsymbol{S_k},\boldsymbol{S_p},\mathbb{C}_i,H_d)\Pr(\mathbb{C}_i)} \quad \text{(A1)}$$

We can now make some simplifications to Eq. (A1), noting:

- We are only considering individuals unrelated to the POI as the other source of DNA in H_p, so $Pr(\boldsymbol{S_{R_j}}|H_p)=\varnothing$.
- We assume it is reasonable to assign equal probabilities for all contributor orders, i.e. $Pr(\mathbb{C}_i)=Pr(\mathbb{C}_j)$.
- The probability of the genotypes of known contributors and POIs given their own genotype is 1 i.e. $Pr(\boldsymbol{S_p}|\boldsymbol{S_p},\mathbb{C}_i,H_p)=1$ and $Pr(\boldsymbol{S_k}|\boldsymbol{S_k},\mathbb{C}_i,H_x)=1$. Note that depending on contributor order these values may be 0 i.e. if the known contributors or POIs are compared to contributor positions where they cannot explain the observed genotype. For simplicity we make the assumption that the list of genotype sets includes only those that S_p and S_k can be compared to contributor positions where they can explain the observed genotype and this leads to a different number of genotype sets being considered under the two propositions, which we distinguish as j and j'.

Using these assumptions and the third law of probability changes Eq. (A1) to:

$$\mathrm{LR}=\frac{\sum_i\sum_j w_j \Pr(\boldsymbol{S_{U_j}}|\boldsymbol{S_p},\boldsymbol{S_k},\mathbb{C}_i,H_p)}{\sum_i\sum_{j'} w_{j'}\Pr(\boldsymbol{S_{R_{j'}}}|\boldsymbol{S_{U_{j'}}},\boldsymbol{S_p},\boldsymbol{S_k},\mathbb{C}_i,H_d)\Pr(\boldsymbol{S_{U_{j'}}}|\boldsymbol{S_p},\boldsymbol{S_k},\mathbb{C}_i,H_d)} \quad \text{(A2)}$$

Note that we could have split $\Pr(\boldsymbol{S_{U_{j'}}},\boldsymbol{S_{R_{j'}}}|\boldsymbol{S_k},\boldsymbol{S_p},\mathbb{C}_i,H_d)$ by $\Pr(\boldsymbol{S_{U_{j'}}}|\boldsymbol{S_{R_{j'}}},\boldsymbol{S_p},\mathbb{C}_i,H_d)\Pr(\boldsymbol{S_{R_{j'}}}|\boldsymbol{S_p},\mathbb{C}_i,H_d)$, which would be as valid as the break up shown in Eq. (3), however doing so would make the calculation more complex as it invokes the consideration of alleles that are IBD between $\boldsymbol{S_{R_{j'}}}$ and $\mathbf{S_p}$ in the calculation of $\Pr(\boldsymbol{S_{U_{j'}}}|\boldsymbol{S_{R_{j'}}},\boldsymbol{S_p},\mathbb{C}_i,H_d)$. This will be explained more fully later on. Eq. (A2) is the general formula for considering a relative as an alternative source of DNA to a profile that can be applied to DNA profiles of any complexity.

We also here note the restricted set of problems we are considering as outlined in Section 2.1, i.e. only those where a single POI is being compared and a single relative in H_d is being considered. In line with the restriction we no longer bold S_R and S_p and change to G_R and G_p to signify they are single genotypes rather than possible vectors containing multiple elements.

C.1. Adding a population genetic model

We start by considering the various elements of Eq. (A2) and explaining their meaning.

$\Pr(\boldsymbol{S_{U_j}}|G_p, \boldsymbol{S_k}, \mathbb{C}_i, H_p)$ – The probability of the unknown individuals', unrelated to the POI, having genotypes that correspond to a specific order of mixture components as designated by contributor order i. This probability is conditioned on having seen the alleles in G_p and $\mathbf{S_k}$. Considering that there is one questioned contributor and K known contributors in an n person mixture we can consider individual genotypes (again using G to signify a single individual's genotype) of unknowns by:

$$\begin{aligned}\Pr(\boldsymbol{S_{U_j}}|G_p, \boldsymbol{S_k}, \mathbb{C}_i, H_p) &= \Pr({}^1G_{U_j}, \ldots, {}^{n-K-1}G_{U_j}|G_p, \boldsymbol{S_k}, \mathbb{C}_i, H_p)\\ &= \Pr({}^1G_{U_j}|{}^2G_{U_j}, \ldots, {}^{n-K-1}G_{U_j}, G_p, \boldsymbol{S_k}, \mathbb{C}_i, H_d)\\ &\quad \times \ldots \times \Pr({}^{n-K-1}G_{U_j}|G_p, \boldsymbol{S_k}, \mathbb{C}_i, H_d)\end{aligned}$$

Where a left superscript designates a single contributor genotype's position within the set.

$\Pr(G_R|\boldsymbol{S_{U_{j'}}}, G_p, \boldsymbol{S_k}, \mathbb{C}_i, H_d)$ – The probability of a relative of the POI having genotype G_R that corresponds to a specific order of mixture components as designated by contributor order i.

$\Pr(\boldsymbol{S_{U_{j'}}}|G_p, \boldsymbol{S_k}, \mathbb{C}_i, H_d)$ – This is the same as $\Pr(\boldsymbol{S_{U_j}}|G_p, \boldsymbol{S_k}, \mathbb{C}_i, H_p)$, although given H_d there is no consideration that the POI is one of the genotypes. This leads to:

$$\begin{aligned}\Pr(\boldsymbol{S_{U_{j'}}}|G_p, \boldsymbol{S_k}, \mathbb{C}_i, H_d) &= \Pr({}^1G_{U_{j'}}|{}^2G_{U_{j'}}, \ldots, {}^{n-K}G_{U_{j'}}, G_p, \boldsymbol{S_k}, \mathbb{C}_i, H_p) \times \ldots\\ &\quad \times \Pr({}^{n-K}G_{U_{j'}}|G_p, \boldsymbol{S_k}, \mathbb{C}_i, H_p)\end{aligned}$$

In the calculation of the probability of G_R consideration must be given to whether there are 0, 1 or 2 alleles that are IBD with the POI. Mathematically these can be considered by:

$$\Pr(G_{R_{j'}}|\boldsymbol{S_{U_{j'}}}, G_p, \boldsymbol{S_k}, \mathbb{C}_i, H_d) = \sum_{x=0}^{2}\Pr(G_{R_{j'}}|\boldsymbol{S_{U_{j'}}}, G_p, \boldsymbol{S_k}, Z_x, \mathbb{C}_i, H_d)\Pr(Z_x) \tag{A3}$$

A further consideration is the evaluation of the probability that a single allele is IBD. Under this scenario either allele in $G_{R_{j'}}$ may be IBD so we split $\Pr(Z_1)$ into $\Pr(Z_{1A})$ and $\Pr(Z_{1B})$, each with probability of ½.

There are a number of genotype-specific scenarios that can now be considered, for example whilst for siblings $\Pr(Z_2) = \frac{1}{4}$, if the siblings do not have matching alleles then clearly the probability that both are IBD is 0. This fact is taken into account already in Eq. (A3), i.e. if $G_{R_{j'}} \neq G_p$ then $\Pr(G_{R_{j'}}|\boldsymbol{S_{U_{j'}}}, G_p, \boldsymbol{S_k}, Z_2, \mathbb{C}_i, H_d) = 0$. However, to allow an automated calculation of $\sum_{x=0}^{2}\Pr(G_{R_{j'}}|\boldsymbol{S_{U_{j'}}}, G_p, \boldsymbol{S_k}, Z_x, \mathbb{C}_i, H_d)\Pr(Z_x)$ using a single equation we extend Eq. (A3) with the addition of an indicator term, α_x:

$$\Pr(G_{R_{j'}}|\boldsymbol{S_{U_{j'}}}, G_p, \boldsymbol{S_k}, \mathbb{C}_i, H_d) = \sum_{x=0}^{2}\Pr(G_{R_{j'}}|\boldsymbol{S_{U_{j'}}}, G_p, \boldsymbol{S_k}, Z_x, \mathbb{C}_i, H_d)\alpha_x\Pr(Z_x)$$

Defining genotypes $G_{R_{j'}} = [a_1, a_2]$ and $G_P = [a_3, a_4]$, and using standard nomenclature so that $a_1 \leq a_2$ and $a_3 \leq a_4$. The indicator terms can be defined by:

$$\alpha_2 = \begin{cases} 1 & a_1 = a_3 \cap a_2 = a_4 \\ 0 & \text{otherwise} \end{cases}$$

$$\alpha_{1A} = \begin{cases} 1 & a_1 = a_3 \cup a_1 = a_4 \\ 0 & \text{otherwise} \end{cases}$$

$$\alpha_{1B} = \begin{cases} 1 & a_2 = a_3 \cup a_2 = a_4 \\ 0 & \text{otherwise} \end{cases}$$

$$\alpha_0 = \begin{cases} 1 & a_1 = a_2 \\ 0 & \text{otherwise} \end{cases}$$

Where α_0 simply accounts for the heterozygosity of $G_{R_{j'}}$. It is important here to remember that $G_{R_{j'}}$ is dependent on contributor order, and so the indicator terms will change depending on the contributor order being assessed. Using the Balding and Nichols sampling formula:

$$\begin{aligned}\Pr(G_{R_{j'}}|\boldsymbol{S_{U_{j'}}}, G_p, \boldsymbol{S_k}, \mathbb{C}_i, H_d) &= \alpha_2\Pr(Z_2)\\ &+ \frac{\alpha_{1B}\Pr(Z_1)}{2}\left\{\frac{x_1\theta + (1-\theta)\,p_{a_1}}{1 + (n_1 - 1)\theta}\right\} + \frac{\alpha_{1A}\Pr(Z_1)}{2}\left\{\frac{x_2\theta + (1-\theta)\,p_{a_2}}{1 + (n_1 - 1)\theta}\right\}\\ &+ \alpha_0\Pr(Z_0)\left\{\frac{\left[x_1\theta + (1-\theta)\,p_{a_1}\right]\left[x_2\theta + (1-\theta)\,p_{a_2}\right]}{[1 + (n_1 - 1)\theta][1 + n_1\theta]}\right\}\end{aligned} \tag{A4}$$

For contributor order i and genotype set j only, where:

- x_1 is the count of allele a_1 seen previously in G_p and other unrelated individuals
- x_2 is the count of allele a_2 seen previously in G_p, other unrelated individuals and a_1 of $G_{R_{j'}}$, when a_1 of $G_{R_{j'}}$ is not IBD with an allele possessed by G_p
- n_1 is the number of alleles seen in total
- p_{a_1} is allele probability for allele a_1
- p_{a_2} is allele probability for allele a_2
- θ is the coancestry coefficient (Fst).

References

[1] S. Turrina, S. Caratti, D. De Leo, Evaluation of PowerPlex® fusion system on samples from forensic casework, Forensic Sci. Int.: Genet. Suppl. Ser. 4 (1) (2013) e210–e211.
[2] D. Taylor, Using continuous DNA interpretation methods to revisit likelihood ratio behaviour, Forensic Sci. Int.: Genet. 11 (2014) 144–153.
[3] D. Taylor, J.-A. Bright, J.S. Buckleton, The interpretation of single source and mixed DNA profiles, Forensic Sci. Int.: Genet. 7 (5) (2013) 516–528.
[4] R. Puch-Solis, et al., Evaluating forensic DNA profiles using peak heights: allowing for multiple donors, allelic dropout and stutters, Forensic Sci. Int.: Genet. 7 (5) (2013) 555–563.
[5] R.G. Cowell, S.L. Lauritzen, J. Mortera, Probabilistic modelling for DNA mixture analysis, Forensic Sci. Int.: Genet. Suppl. Ser. 1 (1) (2008) 640–642.
[6] M.W. Perlin, et al., Validating TrueAllele® dna mixture interpretation, J. Forensic Sci. 56 (6) (2011) 1430–1447.
[7] J. Buckleton, C. Triggs, Relatedness and DNA: are we taking it seriously enough? Forensic Sci. Int. 152 (2005) 115–119.
[8] I.W. Evett, Evaluating DNA profiles in a case where the defence is it was my brother", J. Forensic Sci. Soc. 32 (1) (1992) 5–14.
[9] D. Taylor, et al., The 'factor of two' issue in mixed DNA profiles, J. Theor. Biol. (2014), http://dx.doi.org/10.1016/j.jtbi.2014.08.021.
[10] D.J. Balding, R.A. Nichols, DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands, Forensic Sci. Int. 64 (1994) 125–140.
[11] J.S. Buckleton, C.M. Triggs, Relatedness, in: J.S. Buckleton, C.M. Triggs, S.J. Walsh (Eds.), Forensic DNA Evidence Interpretation, CRC Press, Boca Raton, 2005.
[12] S.J. Walsh, J.S. Buckleton, Autosomal microsatellite allele frequencies for a nationwide dataset from the Australian Caucasian sub-population, Forensic Sci. Int. 168 (2–3) (2007) e47–e50.
[13] J.-A. Bright, et al., Searching mixed DNA profiles directly against profile databases, Forensic Sci. Int.: Genet. 9 (2014) 102–110.

Supplementary material: Below is an expanded version of the LR_2 calculation.

We wish to calculate:

$$\sum_i \sum_{j'} w_{j'} \Pr(G_R \mid \boldsymbol{S}_{U_{j'}}, G_p, \mathbb{C}_i, H_d) \Pr(S_{U_{j'}} \mid \mathbb{C}_i, H_d)$$

Which for the two person scenario in question can be expanded to:

$$\sum_{j'} w_{j'} \Pr(G_R \mid G_{U_{j'}}, G_p, \mathbb{C}_1, H_{d2}) \Pr(G_{U_{j'}} \mid \mathbb{C}_1, H_{d2}) + \sum_{j'} w_{j'} \Pr(G_R \mid G_{U_{j'}}, G_p, \mathbb{C}_2, H_{d2}) \Pr(G_{U_{j'}} \mid \mathbb{C}_2, H_{d2})$$

These terms can be obtained by summing the multiplied element from table 3 so that:

$$\sum_{j'} w_{j'} \Pr(G_R \mid G_{U_{j'}}, G_p, \mathbb{C}_1, H_{d2}) \Pr(G_{U_{j'}} \mid \mathbb{C}_1, H_{d2}) = 0.55\left[\tfrac{1}{4} + \tfrac{1}{4}\left(p_{13} + p_{14}\right) + \tfrac{1}{4} 2 p_{13} p_{14}\right]\left[2 p_{16} p_{17}\right] +$$
$$0.10\left[\tfrac{1}{4} p_{16} + \tfrac{1}{4} 2 p_{13} p_{16}\right]\left[2 p_{14} p_{17}\right] + 0.10\left[\tfrac{1}{4} p_{17} + \tfrac{1}{4} 2 p_{13} p_{17}\right]\left[2 p_{14} p_{16}\right] + 0.10\left[\tfrac{1}{4} p_{16} + \tfrac{1}{4} 2 p_{14} p_{16}\right]\left[2 p_{13} p_{17}\right] +$$
$$0.10\left[\tfrac{1}{4} p_{17} + \tfrac{1}{4} 2 p_{14} p_{17}\right]\left[2 p_{13} p_{16}\right] + 0.05\left[\tfrac{1}{4} 2 p_{16} p_{17}\right]\left[2 p_{13} p_{14}\right]$$

and

$$\sum_{j'} w_{j'} \Pr(G_R \mid G_{U_{j'}}, G_p, \mathbb{C}_2, H_{d2}) \Pr(G_{U_{j'}} \mid \mathbb{C}_2, H_{d2}) = 0.55\left[\tfrac{1}{4} 2 p_{16} p_{17}\right]\left[2 p_{13} p_{14}\right] +$$
$$0.10\left[\tfrac{1}{4} p_{17} + \tfrac{1}{4} 2 p_{14} p_{17}\right]\left[2 p_{13} p_{16}\right] + 0.10\left[\tfrac{1}{4} p_{16} + \tfrac{1}{4} 2 p_{14} p_{16}\right]\left[2 p_{13} p_{17}\right] + 0.10\left[\tfrac{1}{4} p_{17} + \tfrac{1}{4} 2 p_{13} p_{17}\right]\left[2 p_{14} p_{16}\right] +$$
$$0.10\left[\tfrac{1}{4} p_{16} + \tfrac{1}{4} 2 p_{13} p_{16}\right]\left[2 p_{14} p_{17}\right] + 0.05\left[\tfrac{1}{4} + \tfrac{1}{4}\left(p_{13} + p_{14}\right) + \tfrac{1}{4} 2 p_{13} p_{14}\right]\left[2 p_{16} p_{17}\right]$$

Therefore:

$$\sum_i \sum_{j'} w_{j'} \Pr(G_R \mid G_{U_{j'}}, G_p, \mathbb{C}_i, H_d) \Pr(G_{U_{j'}} \mid \mathbb{C}_i, H_d) = \left[2 p_{16} p_{17}\right] \begin{Bmatrix} 0.6\left[\tfrac{1}{4} + \tfrac{1}{4}\left(p_{13} + p_{14}\right) + \tfrac{1}{4} 2 p_{13} p_{14}\right] + \\ 0.2\left[\tfrac{1}{4} p_{14} + \tfrac{1}{4} 2 p_{13} p_{14}\right] + \\ 0.2\left[\tfrac{1}{4} p_{14} + \tfrac{1}{4} 2 p_{13} p_{14}\right] + \\ 0.2\left[\tfrac{1}{4} p_{13} + \tfrac{1}{4} 2 p_{14} p_{13}\right] + \\ 0.2\left[\tfrac{1}{4} p_{13} + \tfrac{1}{4} 2 p_{14} p_{13}\right] + \\ 0.6\left[\tfrac{1}{4} 2 p_{13} p_{14}\right] \end{Bmatrix}$$

and the *LR* can be calculated by:

$$LR_2 = \frac{1.2 \times p_{16} p_{17}}{[2 p_{16} p_{17}] \left\{ \begin{array}{l} 0.60\left[\frac{1}{4} + \frac{1}{4}\left(p_{13} + p_{14}\right) + \frac{1}{4} 2 p_{13} p_{14}\right] + 0.20\left[\frac{1}{4} p_{14} + \frac{1}{4} 2 p_{13} p_{14}\right] + \\ 0.20\left[\frac{1}{4} p_{14} + \frac{1}{4} 2 p_{13} p_{14}\right] + 0.20\left[\frac{1}{4} p_{13} + \frac{1}{4} 2 p_{14} p_{13}\right] + \\ 0.20\left[\frac{1}{4} p_{13} + \frac{1}{4} 2 p_{14} p_{13}\right] + 0.60\left[\frac{1}{4} 2 p_{13} p_{14}\right] \end{array} \right\}}$$

$$LR_2 = \frac{1.2}{0.30 + 0.50 p_{13} + 0.50 p_{14} + 2 p_{13} p_{14}}$$

$$LR_2 = \frac{12}{3 + 5\left(p_{13} + p_{14}\right) + 20 p_{13} p_{14}}$$

Chapter 4: Calibrating the model to specific laboratory performance

STRmix™ incorporates models that are used to describe DNA profile behaviour. These include models for stochastic events such as peak height variability and inter-locus balances. When two different processes are used to generate DNA profiles, the profiles will naturally exhibit different behaviours (as in they will still be described by the same models, but the parameter values within the models will vary). There are two ways of dealing with this fact:

1) Have these aspects of uncertainty as models within the MCMC, whose parameter values are guided from the data being analysed
2) Carry out calibration testing and fix some parameters for the type of data being analysed

STRmix™ uses the second of these methods. There are other fully continuous systems in existence that work by the first. Anecdotal feedback is that the runtime is greatly increased and often there is not enough information in the data provided to inform some parameters, meaning the results produced are not always intuitive.

The standard implementation of STRmix™ in a forensic laboratory is to create a calibration set of profiles (typically 100 or more) and run them through a calibration program that aligns STRmix™ models to laboratory performance. If a laboratory has more than one workflow (i.e. they have situations which mean that DNA samples may be profiled by one of multiple different DNA profiling systems) then a calibration is required for each. When STRmix™ was introduced questions arose as to how different the processes needed to be in order to justify a new calibration set. These questions were initially quite prevalent as early versions of STRmix™ had a peak height variability model that utilised a single constant, which aligned it to the amount of variability seen in data produced by the laboratory. In 2014 (approximately 2 years after initial release of STRmix™) the peak height variability model was updated so that the constant was now a parameter within the MCMC, that had a prior distribution, produced from calibration. This new model architecture allowed the tolerance of the system to peak imbalances to shift slightly (and in accordance with the expectations of data produced by the laboratory) depending on the data being analysed. This largely addressed questions of whether many calibration sets were required for micro-variations in laboratory process. As long as the calibration set was created in a way that covered many of those micro-variants of laboratory process then the prior distribution for peak height variability would be applicable for a wide range of data produced and could be fine-tuned to the specific profiles as required during the MCMC.

Still the question remained '*How much difference requires recalibration*?'. Examples may be:

- If laboratories had a piece of equipment changed
- From above, consideration of whether the new equipment was of the same model as the old
- If their equipment was serviced
- If a component was replaced

These types of questions led to the work presented in this section. Also, within this work, the component-wise MCMC process utilised in the calibration tool (called 'Model Maker' that made up part of STRmix™) is described.

Manuscript: Factors affecting peak height variability for short tandem repeat data. D Taylor, J Buckleton, JA Bright. (2016) Forensic Science International: Genetics 21, 126-133 – *Cited 2 times*

Statement of novelty: This work provides a description of how the data produced by a laboratory can be used to calibrate STRmix™ for its specific performance. This work extended the MCMC theory given in section 2.6 to achieve the desired outcome.

My contribution: Main theorist and author of the work. Sole programmer of simulations.

Research Design / Data Collection / Writing and Editing = 60% / 40% / 70%

Additional comments:

Forensic Science International: Genetics 21 (2016) 126–133

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

Research paper

Factors affecting peak height variability for short tandem repeat data

Duncan Taylor[a,b,*], John Buckleton[c,d], Jo-Anne Bright[c]

[a] Forensic Science South Australia, 21 Divett Place, Adelaide, SA 5000, Australia
[b] School of Biological Sciences, Flinders University, GPO Box 2100, Adelaide, SA 5001, Australia
[c] ESR, Private Bag 92021, Auckland 1142, New Zealand
[d] National Institute of Standards and Technology, 100 Bureau Drive, MS 8980 and 8314, Gaithersburg, MD 20899, United States of America

ARTICLE INFO

Article history:
Received 11 August 2015
Received in revised form 26 November 2015
Accepted 16 December 2015
Available online 19 December 2015

Keywords:
Peak height variability
Stochastic effects
STRmix
STR

ABSTRACT

In forensic DNA analysis a DNA extract is amplified using polymerase chain reaction (PCR), separated using capillary electrophoresis and the resulting DNA products are detected using fluorescence. Sampling variation occurs when the DNA molecules are aliquotted during the PCR setup stage and this translates to variability in peak heights in the resultant electropherogram or between electropherograms generated from a DNA extract. Beyond the variability caused by sampling variation it has been observed that there are factors in generating the DNA profile that can contribute to the magnitude of variability observed, most notably the number of PCR cycles. In this study we investigate a number of factors in the generation of a DNA profile to determine which contribute to levels of peak height variability.

1. Introduction

In forensic DNA analysis a DNA extract is amplified using polymerase chain reaction (PCR), separated using capillary electrophoresis and the resulting DNA products are detected using fluorescence. The result is a trace of fluorescence against time termed an electropherogram (epg). The largest peaks in the epg are usually interpreted as signals from amplified DNA product. The height of these peaks is approximately linearly proportional to the starting number of DNA molecules, termed template [1,2]. However if replicate aliquots are taken from the same extract the resulting peaks will not all be of the same height. We could consider these peaks to be varying around a mean value and this variability could be quantified by the variance.

There have been numerous studies of the variability of peak heights in short tandem repeat (STR) profiles (for example see Refs. [1,3]). The variability in peak heights has led to a variety of interpretation guidelines that are used to accept or reject potential genotypic explanations (or genotypes) for some observed data in an epg. More recently the variability in peak heights has been modelled with distributions for use in continuous interpretation systems [4–7]. The question arises whether all DNA profiles produced within a laboratory have the same amount of peak height variability. It may be that that the presence of factors such as DNA amount, the presence of inhibitors, micro-variation in amplification or electrophoretic conditions, means that all profiles are not created equally with respect to the amount of peak height variability present. There are also questions that arise regarding larger scale differences that can cause peak height variability between epgs such as those having been amplified using different profiling kits, using different thermocycler instruments, or run on different genetic analysers, either of the same or different models.

Classically peak height variability is measured by looking at heterozygous balance (Hb) [3,8,9]. In this way the two peaks of a heterozygous pair can act as an internal self-calibrated pair of peaks for the template at each locus in each profile. Given that the expected variance in log(Hb) is twice the expected variance of the log of the expected individual peak heights at that locus (see Ref. [10] for a proof) we can also use the latter to assess peak height variability across a profile.

Allelic peaks are not the only peaks in an epg. Other peaks, often termed artefactual peaks, are present. The largest and most common artefact is backward stutter. These peaks would also vary in height if replicate extracts were taken.

In the model of Taylor et al. [6] the height of each allele and stutter peak is considered to vary about some expected value modelled from the profile. If the expected value is obtained correctly then the difference from expectation of each peak is independent. The model used to compare an observed peak height

* Corresponding author at: Forensic Science South Australia, 21 Divett Place, Adelaide, SA 5000, Australia. Fax: +61 8 8226 7777.
E-mail address: Duncan.Taylor@sa.gov.au (D. Taylor).

http://dx.doi.org/10.1016/j.fsigen.2015.12.009
1872-4973/

(O) and an expected peak height (E) is given by:

$$\log_{10}\left(\frac{O_a}{E_a}\right) \sim \mathrm{N}\left(0, \frac{c^2}{E_a}\right) \text{for allelic peak } a \text{ and}$$
$$\log_{10}\left(\frac{O_{a-1}}{E_{a-1}}\right) \sim \mathrm{N}\left(0, \frac{k^2}{O_a}\right) \text{for stutter peak a} - 1 \quad (1)$$

where O_a and $O_{a\text{-}1}$ are the observed height of the allele and stutter peaks, respectively, and E_a and $E_{a\text{-}1}$ are the expected height of the allele and stutter peaks. c^2 and k^2 are random variables which are determined based on empirical data and describe the variability within a dataset. The values of c^2 and k^2 for a dataset are determined per locus.

In this paper, we use a Markov chain Monte Carlo (MCMC) system to analyse sets of laboratory data of known origin in order to determine the distribution of peak height variability. This is done for a number of datasets, which vary in the mechanical aspects of profile generation, including multiplex used, the cycle number, and the genetic analyser.

We also compare the distributions obtained from laboratory data with data simulated with known values of c^2 and k^2 within the peak height variability models. Doing so allows conclusions to be drawn as to whether c^2 and k^2 act like fixed unknown values that produce an observed distribution of peak height imbalances or whether individual profiles possess differing levels of peak height variability, whose distribution cannot be explained by fixed c^2 and k^2 values (suggesting that we have not yet identified all contributing factors to peak height variability). Any difference has implications for interpretation guidelines, if they are based on factors affected by peak height variability, such as heterozygous balance or dropout.

2. Method

Methods are in four sections. The first section is a description of the statistical process by which peak height variability is assessed. The second section describes the experiments that investigate factors that affect peak height variability. The third section investigates whether profiles produced from a laboratory could all have arisen from the same underlying peak variability or whether micro-variations give rise to some data being more variable than other. The fourth section investigates aspects such as how the source of DNA or the complexity of the profile contributes to peak height variability.

2.1. Section 1: The statistical model

An epg can be considered as a series of observed peaks, each with an allelic designation, a height and a molecular weight. Models have been developed that describe DNA profile behaviour, for example models that link DNA amount and degradation [11] to fluorescence, stuttering during PCR or locus amplification efficiencies [7]. Given these models, values for parameters within them can be trialled, along with genotypes, to build up an expected profile, comprising a number of expected peak heights of various allelic designations at specific molecular weights. Any difference between observed and expected heights must then be explained by peak height variance models (which may include instances of drop-out and drop-in). The difference between each individual pairing of an observed and expected peak height can be modelled by Eq. (1), which contains variable c^2. In Taylor et al. [6] the value of c^2 was set at a point value. However we now consider c^2 to be a random variable within the model, which itself has a prior gamma distribution that can be obtained from laboratory calibration data.

As described in Ref. [6], MCMC can be used to provide a weight representing the probability density of the profile given a postulated genotype set. Fig. 1 shows this diagrammatically. A typical analysis would provide the evidence (in the form of DNA profile(s)), utilise pre-defined models (for each of the parameters being considered, such as degradation, template amounts, locus amplification efficiencies, peak height variability, etc.) and use MCMC to elucidate the weights (in this case a normalised posterior probability density for a proposed genotype set). Although mathematically unnecessary we normalise the weights at each locus so that they range from zero (indicating that the observed data cannot be obtained from the proposed genotype) to one (indicating that it is certain the observed data came from the proposed genotype). This is done so the weights can provide an intuitively helpful diagnostic for analysts.

The same system shown in Fig. 1 can be used to inform an analyst about parameters within the models, if provided with the other points of the triangle. Instead of providing evidence profiles and model parameters to elucidate the weights, the system can be provided evidence profiles and weights (or more specifically the known genotypes of individuals) to elucidate the parameter values within the models (particularly variance variables c^2 and k^2). Such a system still requires some architecture for the biological models being used (e.g. that given in Eq. (1)), however specifics, such as the distribution of c^2 in the modelling of peak height variance, can be determined. The simplest weights to provide are those for single sourced profiles of known origin. In these cases weights will all be one for the genotype that corresponds to the known source.

To determine the peak height variability for profiles generated at a laboratory using a certain process requires multiple profiles so that the range of peak height variabilities encountered is captured. This is particularly true if we wish to consider the premise that some profiles possess more variability in peak heights than others. We analyse a dataset comprising multiple single sourced profiles, representing a range of peak heights, and determine the distributions of the peak height variability variables (k^2 for stutter peaks and c^2 for allelic peaks) for the entire dataset simultaneously. We refer to the distributions as hyper-distributions as they are distributions for variables that are used in further distributions. Eq. (1) shows how the log of the observed over the expected peak heights are modelled using a normal distribution that uses c^2 and k^2 variables within the variance. The variables c^2 and k^2 themselves have prior distributions and can be thought of as distributions within distributions. We therefore refer to the prior distributions as hyper-distributions and the parameters for a hyper-distribution as 'hyper-parameters'. This is carried out using component-wise MCMC by splitting the 'mass' parameters $\boldsymbol{M}$ (which we explain below), and the variance variable hyper-distributions into different components and varying them separately. The variance variables are modelled by gamma distributions, $\Gamma(\alpha_x, \beta_x)$, where x is a subscript used to highlight that the c^2 and k^2 hyper-distributions have different hyper-parameters and the probability density

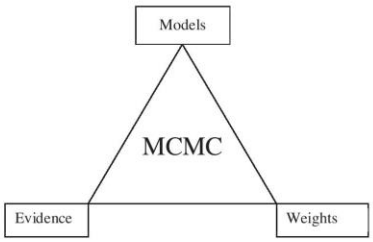


Fig. 1. Diagrammatic representation of the MCMC process used in DNA profile analysis.

Table 1
Experimental datasets analysed for Section 2.

Dataset	DNA profiling kit	PCR cycles	Electrophoretic instrument	Thermocycler instrument	Dataset size	PCR volume
1	Profiler Plus™	28	3130xl (instrument 2)	9700 (numerous)	342	25 μL
2	Profiler Plus™	28	3130xl (instrument 2)	9700 (instrument 5)	93	25 μL
3	Profiler Plus™	28	3130xl (instrument 2)	9700 (instrument 8)	47	25 μL
4	Profiler Plus™	26	3130xl (instrument 1)	9700 (numerous)	233	25 μL
5	Profiler Plus™	26	3130xl (instrument 1)	9700 (instrument 6)	134	25 μL
6	Profiler Plus™	26	3130xl (instrument 2)	9700 (instrument 6)	49	25 μL
7	GlobalFiler™	29	3130xl (instrument 1)	9700 (numerous)	338	25 μL
8	PowerPlex 21™	29	3130xl (instrument 1)	9700 (numerous)	70	12.5 μL
9	PowerPlex 21™	29	3130xl (instrument 1)	9700 (numerous)	89	25 μL
10	GlobalFiler™	29	3130xl (instrument 3)	9700 (numerous)	129	25 μL
11	GlobalFiler™	29	3500	9700 (numerous)	85	25 μL
12	Identifiler™	28	3130xl (instrument 3)	9700 (numerous)	91	25 μL
13	MiniFiler™	30	3130xl (instrument 3)	9700 (numerous)	61	50 μL
14	SGMPlus™	34	3130xl (instrument 3)	9700 (numerous)	106	50 μL

function at point i is given by:

$$f(i|\alpha,\beta) = \frac{i^{\alpha-1}e^{-i/\beta}}{\Gamma(\alpha)\beta^{\alpha}}$$

This distribution was chosen as it has properties that align with the expectations of the variance variables i.e. bounded at the lower end by zero, at the upper end by infinity, and likely to be asymmetrical. The inverse proportionality of the stutter peak height variance variable is based on observed parent peak height, O_a, rather than its own expected peak height.

The mass parameters are; a DNA amount and a degradation, for each profile, and a locus amplification efficiency, for each locus in each profile. Separate variance variable values are sampled for each locus in each profile from the hyper-distributions as part of $\boldsymbol{M}$.

The c^2 and k^2 gamma hyper-distributions parameters (α_x and β_x for each of the two gamma hyper-distributions) are collectively referred to as vector $\boldsymbol{V}$. We also have an exponential hyper-distribution for locus specific amplification efficiency within $\boldsymbol{V}$, with probability density function:

$$f(i|\lambda) = \lambda e^{-\lambda i} \qquad i \geq 0$$

Parameter values within $\boldsymbol{V}$ apply to all profiles collectively as a group. The mass parameters for the C profiles in the dataset, $\boldsymbol{M}_1 \ldots \boldsymbol{M}_C$, are varied one at a time (separately from $\boldsymbol{V}$ and from each other). The analysis is then carried out by repeating loops of:

1) Varying $\boldsymbol{M}_1 \ldots \boldsymbol{M}_C$ while holding $\boldsymbol{V}$ constant
2) Varying $\boldsymbol{V}$ while holding $\boldsymbol{M}_1 \ldots \boldsymbol{M}_C$ constant

until all values within both vectors have converged (diagrammatically seen in Fig. 2). The MCMC used is a random walk component-wise MCMC using a Metropolis–Hastings sampler. We use this as it has previously been successfully used in DNA profile mixture deconvolution by Curran [12] and has been shown more extensively to perform well on a range of DNA profiles of varying

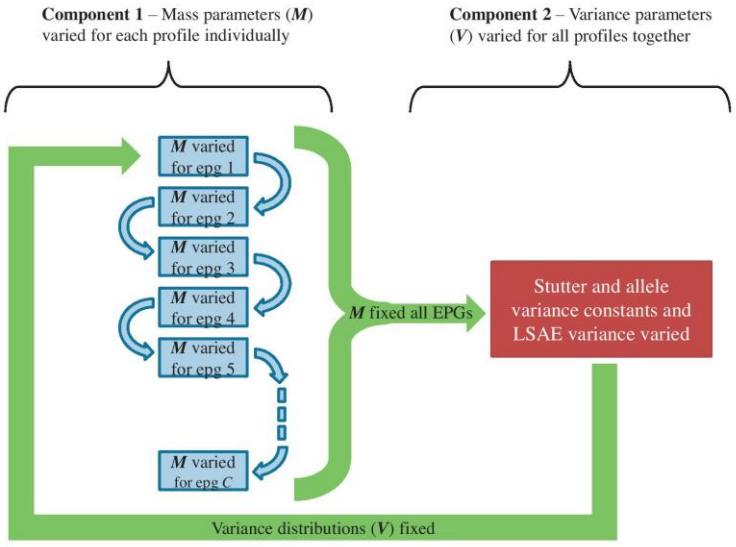


Fig. 2. Component-wise nature of MCMC process to determine $\boldsymbol{V}$.

Table 2
Allele variance variable hyper-distributions for datasets outlined in Table 1.

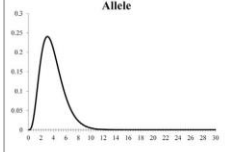
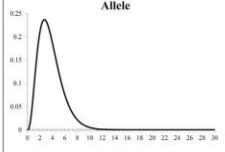
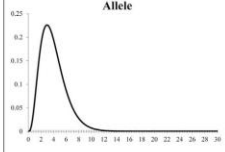
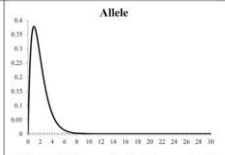
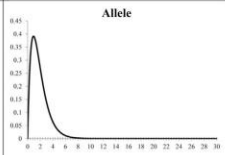
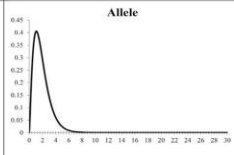
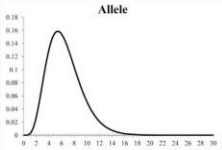
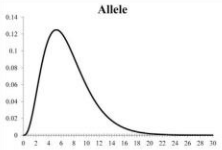
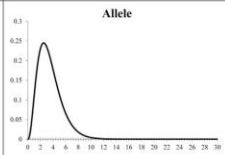
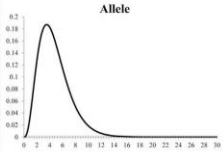
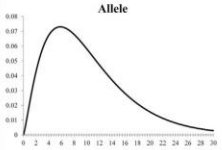
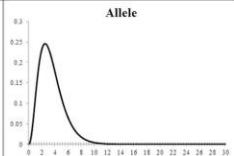
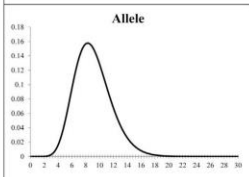
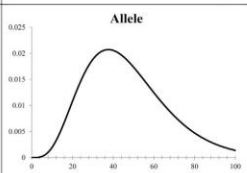
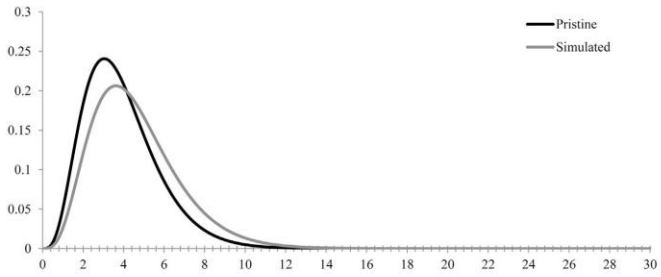
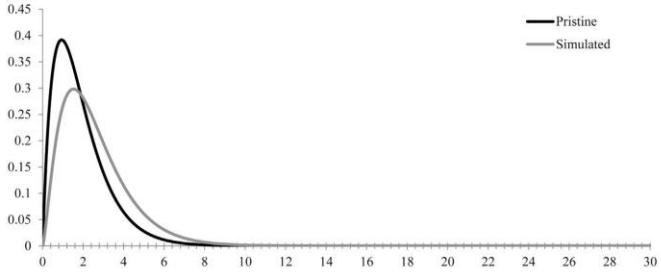
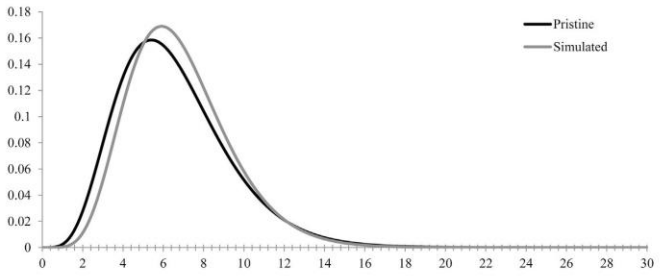
Set	1	2	3
Profiler Plus 28 cycles25μL			
	$\Gamma(4.53,0.862)$, mode 3.04, mean 3.9	$\Gamma(3.769,0.983)$, mode 2.72, mean 3.71	$\Gamma(3.962,0.997)$, mode 2.95, mean 3.95
Set	**4**	**5**	**6**
Profiler Plus 26 cycles 25μL			
	$\Gamma(2.006,0.971)$, mode 0.98, mean 1.95	$\Gamma(1.994,0.942)$, mode 0.94, mean 1.88	$\Gamma(2.339,0.801)$, mode 1.07, mean 1.87
Set	**7**	**8**	**9**
29 cycles	GlobalFiler 25 μL	PowerPlex 21 12.5 μL	PowerPlex 21 25 μL
			
	$\Gamma(5.78,1.131)$, mode 5.41, mean 6.54	$\Gamma(3.84,1.838)$, mode 5.22, mean 7.06	$\Gamma(3.54,0.99)$, mode 2.51, mean 3.50
Set	**10**	**11**	**12**
	GlobalFiler™ - 29 cycles – 3130xl 25μL	GlobalFiler™ - 29 cycles – 3500 25μL	Identifiler™ - 28 cycles 25 μL
			
	$\Gamma(3.907,1.215)$, mode 3.53, mean 4.75	$\Gamma(2.295,4.507)$, mode 5.84, mean 10.34	
Set	**13**	**14**	
	MiniFiler™ - 30 cycles – 50 μL	SGMPlus - 34 cycles – 50 μL	$\Gamma(3.569,0.982)$, mode 2.52, mean 3.50
			
	$\Gamma(11.906,0.760)$, mode 8.29, mean 9.05	$\Gamma(4.983,9.449)$, mode 37.64, mean 47.08	

Table 3
Allelic variance variable distribution for pristine produced datasets and the same distributions shown for datasets simulated to align with mode of pristine results.

Set	Laboratory c^2 hyper-distribution
1	 Pristine - $\Gamma(4.53,0.862)$, mode 3.04, mean 3.9 Simulated - $\Gamma(4.672,0.987)$, mode 3.62, mean 4.61
5	 Pristine - $\Gamma(1.994,0.942)$, mode 0.94, mean 1.88 Simulated - $\Gamma(2.457,1.048)$, mode 1.53, mean 2.57
7	 Pristine - $\Gamma(5.78,1.131)$, mode 5.41, mean 6.54 Simulated - $\Gamma(7.44,0.918)$, mode 5.91, mean 6.83

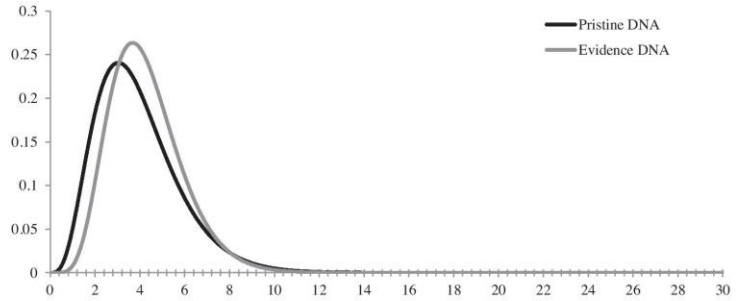


Fig. 3. Peak height variance variable distribution for pristine DNA (Dataset 1) and casework evidence profiles produced under the same conditions.

quality and complexity [6,13].

The result is two gamma hyper-distributions that describe the multiple individual locus/profile variance variables' values for stutter and allele peaks, and one exponential hyper-distribution that describes the multiple individual locus/profile amplification efficiencies and C lots of mass parameters, one for each profile.

2.2. Section 2

We calculate the stutter and allele hyper-distributions for a number of datasets varying:

1) DNA profiling kit
2) Number of PCR cycles
3) Genetic analysers of different models
4) Different genetic analysers of the same model
5) Different thermocycler instruments of the same model
6) Different dataset size
7) Different PCR reaction volumes

Note that although it is already known that the number of PCR cycles and different models of genetic analysers have an effect on peak height variability, the method we outline provides a quantification of these differences. Table 1 gives the breakdown of datasets analysed.

2.3. Section 3

For this experiment we require datasets where all profiles have the same known point value for c^2 and for k^2. There is no way that data can be produced under these conditions within a laboratory. To overcome this limitation a program was written that simulates profiles from a distribution with a known underlying peak height variability and random DNA amounts, degradation values, and locus amplification efficiencies using the same models as described in Ref. [6]. Three datasets of the same sample size as datasets 1, 5 and 7 were simulated to have variance variable values that mimic the modes of the hyper-distributions for those dataset. These datasets were chosen as they represent a range of c^2 and k^2 distributions encountered in standard casework.

If data that is produced in a laboratory has a range of peak height variances then we would expect the density distribution of the hyper-distributions to be spread over a larger range of c^2 values than a dataset simulated to have that exact level of peak height variability.

2.4. Section 4

In this section we address two common questions:

1) Is pristine DNA representative of DNA results obtained in typical casework samples? We define 'pristine' data as laboratory data from uninhibited and undegraded reference DNA.
2) Does the number of contributors to a DNA profile affect peak height variability i.e. are mixtures more variable than single source profiles?

To answer the first question, single source evidence profiles obtained in casework at FSSA over a two week period were compiled into a dataset ($N = 136$) and analysed as described in

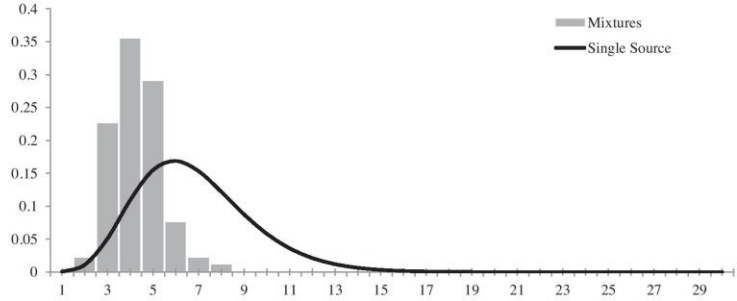


Fig. 4. Peak height variance variable distribution for single source DNA (Dataset 7) and posterior mean values of c^2 for mixed DNA profiles produced under the same conditions.

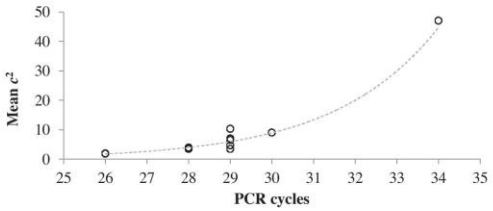


Fig. 5. Relationship between PCR cycles and c^2.

Section 1. These samples have been run under the same conditions as dataset 1 in Table 1. These samples were chosen such that a reasonable inference as to the true contributing genotype could be made. If casework samples have greater peak height variability than pristine samples then we would expect a shift in the density distribution to higher values.

An experiment to answer the second question is more difficult to design. We used the peak height variance variable distribution for dataset 7 as a prior distribution to analyse 93 mixed DNA profiles produced under the same conditions (refer.Table 1) allowing the variance variable to vary as a parameter throughout the analysis. The 93 mixed DNA samples are those described in Ref. [13]. All mixtures were analysed assuming all known contributors to the DNA profile so that any peak height variability had to be described by stochastic variation rather than ambiguity in the genotypes of contributors. Note that there will still remain some ambiguity as to the amount of DNA that each contributor, the level of degradation that each contributor exhibits and the level of amplification efficiency that exists at each locus. The effect of this is discussed later.

If mixtures generally have a higher peak height variability than single sourced profiles then we would expect the distribution of the posterior mean values of c^2 for the 93 mixed GlobalFiler™ samples to have more mass at higher variance values than expected from the prior alone.

3. Results

3.1. Section 2

The result of the datasets described in Table 1 are given in Table 2. We show only the results of the allelic variable (c^2) hyper-distributions as they are most influential to an analyst's typical interpretation of DNA profiles. Hyper-distributions for the stutter variable (k^2) and for the locus amplification efficiencies were determined but are not shown.

3.2. Section 3

Table 3 shows the allelic variance variable distributions for selected datasets shown in Table 2 and the equivalent distributions for datasets simulated to c^2 and k^2 values that align with the mean of the 'real' datasets and a locus amplification efficiency variance that aligns with the mean of the exponential hyper-distribution.

3.3. Section 4

Fig. 3 shows the peak height variance variable distributions for dataset 1 (shown in black as 'pristine' data) and 136 single sourced evidence profiles (shown as grey 'evidence' data) that ranged in sample type and strength.

Fig. 4 shows the posterior distribution (in black) for the peak height variance variable used (based on dataset 7) in the analysis of 93 mixed DNA profiles that have properties that align with dataset 7 (i.e. kit, cycle number, etc). The histogram in Fig. 4 shows the means of the posterior value for c^2 for each of the analysed mixed DNA samples.

4. Discussion

Multiple regression of the mean of the gamma distributions produced in Section 2 suggests that the largest factor is PCR cycle number. The increase in mean c^2 values appears exponential with respect to cycle number (even with the noise present from other variables), as seen in Fig. 5.

There was also an increase in mean c^2 (of approximately 3.6) going from a reaction volume of 25 μL to 12.5 μL (although this is based on only a single comparison between datasets 8 and 9). Kit choice was also a significant factor, although there was insufficient overlap with other variables to explore this adequately. A small effect was present for electrophoresis instrument of the same model and this was not significant on this test, but a much larger effect was seen transitioning from 3130xl to 3500 (again based on

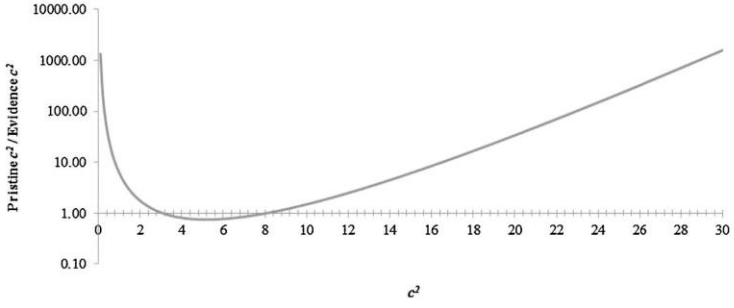


Fig. 6. Ratio of c^2 values from pristine and evidence profile shown in Fig. 3.

only a single dataset comparison), which is expected given the difference in fluorescence scaled used between the two instruments.

We use the means of the c^2 distributions as a point of comparison as it is difficult to directly compare distributions in some way that allows a simple conclusion to be drawn such as "*process A exhibits higher peak height variability than process B*". A comparison of gamma distributions could tell us that they are different, but not how they are different. For example if we look at the results of pristine compared to casework DNA profiles shown in Fig. 3, then it may be tempting to state that visually it appears pristine DNA exhibits slightly lower peak height variability than evidence profiles as the mode of the distribution is lower. This only tells part of the story. Fig. 6 shows the ratio of the density of the two gamma distributions from Fig. 3. Indeed we can see that at the lower levels of c^2 the pristine data have higher density than evidence data, but notice that this also occurs for high levels of c^2. This may suggest that pristine data show a greater spread of profile variances. Fig. 3 equally shows that evidence samples may have slightly higher peak height variability than pristine profiles for the majority of samples that will be encountered.

A reasonable argument could be made that the right hand tail of Fig. 6 is not relevant as it is in an unrealistic peak height variability space. Indeed the highest evidence profile c^2 value observed was 11.6 and the gamma curve beyond this is entirely extrapolation. Our summaries of whether a difference exists are therefore based on the comparison of means and the level of overlap between credible intervals for the distributions.

The pristine laboratory data shows no excess of variability over the profiles simulated from an exact point value for underlying variance (see Section 3). This suggests that single values for the variance variables for allele and stutter may suffice to model these profiles.

There may be sporadic evidence DNA profiles that (perhaps due to interactions between the PCR process and some foreign material) have higher peak height variability than expected from calibration data. We have been anecdotally informed that this is the case by caseworking scientists, although did not view such an occurrence within this study. In general we found that pristine DNA has approximately the same peak height variability as casework samples. This result is consistent with earlier work [14] addressing the same issue. This indicates that there are no issues validating systems using pristine DNA to develop and refine DNA profile behaviour models.

Mixed DNA profiles are likely to be no more variable in peak height (and perhaps less so) than single source DNA profiles. There are two points to consider here:

1) Mixed DNA profiles have additional dimensions (in this case an extra DNA amount and an extra degradation amount for each contributor) that can vary to best explain the observed data, and hence reduce the apparent peak height variability compared to single sourced profiles.
2) The method of mixture analysis has a variance variable that is the posterior mean across the whole profile, rather than a locus at a time. The effect is an 'averaging' across the profile, which may mask a single outlying high stochastic occurrence.

Even given these two points the results shown in Fig. 4 do not indicate an increase in peak height variability in mixtures compared to single source profiles.

Datasets 7 and 10 (Section 2) are equivalent in terms of the factors noted in this study, but show some variation in the distribution of the peak height variance variable. While these two datasets were set-up and run using the same conditions they were done so in different laboratories. The differences observed, whilst not marked, suggest that there are factors not captured in this study that could be having an effect on peak height variability. Possibilities include CE laser sensitivity, CE capillary age, laboratory plastic ware used, age of reagents, PCR setup conditions, or any number of other variants. This suggests that ideal practice would be for each laboratory to diagnose their own performance prior to any analyses or interpretations.

5. Conclusion

Peak variability appears to increase with cycle number and with reduced reaction volume. Additionally, profiling kits and CE models have an effect on peak height variability. There may be a small increase in variability for evidence DNA over pristine DNA, although if this trend is present it is not significant in our findings.

There is no evidence for an increase in variability of mixed samples over single source or that different CE or thermocycler instruments of the same model have a noticeable effect on peak height variability. However there must be other variables, not examined in this work, that have an effect on peak height variability.

Acknowledgements

This work was supported in part by grant 2014-DN-BX-K028 from the National Institute of Justice. Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the U.S. Department of Justice or Commerce. Certain commercial equipment, instruments, or materials (or suppliers, or software) are identified in this paper to foster understanding. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

References

[1] P. Gill, J. Curran, K. Elliot, A graphical simulation model of the entire DNA process associated with the analysis of short tandem repeat loci, Nucleic Acids Res. 33 (2005) 632–643.
[2] M.D. Timken, S.B. Klein, M.R. Buoncristiani, Stochastic sampling effects in STR typing: implications for analysis and interpretation, Forensic Sci. Int. Genet. 11 (2014) 195–204.
[3] J.-A. Bright, J. Turkington, J. Buckleton, Examination of the variability in mixed DNA profile parameters for the Identifiler(TM) multiplex, Forensic Sci. Int. Genet. 4 (2010) 111–114.
[4] M.W. Perlin, M.M. Legler, C.E. Spencer, J.L. Smith, W.P. Allan, J.L. Belrose, et al., Validating TrueAllele® DNA mixture interpretation, J. Forensic Sci. 56 (2011) 1430–1447.
[5] R. Puch-Solis, L. Rodgers, A. Mazumder, S. Pope, I. Evett, J. Curran, et al., Evaluating forensic DNA profiles using peak heights, allowing for multiple donors, allelic dropout and stutters, Forensic Sci. Int. Genet. 7 (2013) 555–563.
[6] D. Taylor, J.-A. Bright, J. Buckleton, The interpretation of single source and mixed DNA profiles, Forensic Sci. Int.Genet. 7 (2013) 516–528.
[7] J.-A. Bright, D. Taylor, J.M. Curran, J.S. Buckleton, Developing allelic and stutter peak height models for a continuous method of DNA interpretation, Forensic Sci. Int. Genet. 7 (2013) 296–304.
[8] J.-A. Bright, E. Huizing, L. Melia, J. Buckleton, Determination of the variables affecting mixed MiniFiler (TM) DNA profiles, Forensic Sci. Int. Genet. 5 (2011) 381–385.
[9] H. Kelly, J.-A. Bright, J.M. Curran, J. Buckleton, Modelling heterozygote balance in forensic DNA profiles, Forensic Sci. Int. Genet. 6 (2012) 729-–34.
[10] D. Taylor, J.-A. Bright, C. McGoven, C. Hefford, T. Kalafut, J. Buckleton, Validating multiplexes for use in conjunction with modern interpretation strategies, Forensic Sci. Int. Genet. 20 (2016) 6–19. http://dx.doi.org/.</doc_ref&gt10.1016/j.fsigen.2015.09.011.
[11] J.-A. Bright, D. Taylor, J.M. C, J.S. Buckleton, Degradation of forensic DNA profiles, Aust. J. Forensic Sci. 45 (2013) 445–449.
[12] J. Curran, A MCMC method for resolving two person mixtures, Sci. Justice 48 (2008) 168–177.
[13] D. Taylor, Using continuous DNA interpretation methods to revisit likelihood ratio behaviour, Forensic Sci. Int. Genet. 11 (2014) 144–153.
[14] J.-A. Bright, K. McManus, S. Harbison, P. Gill, J. Buckleton, A comparison of stochastic variation in mixed and unmixed casework and synthetic samples, Forensic Sci. Int. Genet. 6 (2012) 180–184.

4 – Clarification

Explanation of the model, specifically framing it in the context of a hierarchical Bayesian model

The model used in STRmix is one where we seek the likelihood of genotype sets, given the observed data, $\int \sum_j p(\mathbf{O} \mid \mathbf{S_j}, \boldsymbol{M}) \Pr(\mathbf{S_j} \mid, H_1) \Pr(\boldsymbol{M}) d\boldsymbol{M}$. Within this integral the **M** term represents parameters:

- Template DNA amount for each contributor (*n*), which has prior $t_n \sim U[0, T]$ (where T represents the upper limit on template amount before a DNA profile will no longer be analysed and is termed a saturation level)
- Degradation for each contributor, which has prior $d_n \sim U[0, D]$ (where *D* represents a level of degradation above which profiles will generally be considered too low quality and will not be analysed)
- A PCR replicate efficiency term for each PCR replicate (*y*), which has prior $R_y \sim U[0, \infty]$ (note that in practise, if an analysis was carried out and a replicate amplification efficiency obtained beyond the approximate bounds [0.1,10] it would be considered that one of the replicates is likely to have been the subject of an amplification error and should not be included in the analysis)
- An amplification efficiency term for each locus (*l*), which has prior $A^l \sim LN\left(0, \xi^2 \sigma^2\right)$ (where $\xi = \ln(10)$ is used to transform between logs in base 10 and base *e* and σ^2 is constant, determined by laboratory calibration)
- A peak height variability parameter for each fluorescence type (*i*), which has prior $c^i \sim \Gamma\left(\alpha^i, \beta^i\right)$ (determined by laboratory calibration, which I discuss below)

Knowing the values of the parameters in **M** allows the calculation of Total Allelic Product (*T*), the total amount of fluorescence expected resulting from an allele in a DNA extract, which will ultimately get broken into components of fluorescence in an allelic position and its stutter positions on the electropherogram (EPG). Calculation of *T*, for a combination of contributor, kit, locus, replicate and allele, is achieved formulaically by:

$$T^l_{a,n,y} = t_n \times A^l \times R_y \times X^l_n \times e^{-d_n\left(m^l_a - offset\right)}$$

The X^l_n term in equation 1 represents a 'dose' and takes values of 1 or 2. The dose considers that if contributor *n* is homozygous for allele *a* at locus *l* ($X^l_n = 2$), then the expected value for *T* will be twice as high than if allele *a* was one in a heterozygous pair. The offset marks the molecular weight at which degradation starts to be applied, i.e. at the offset (and technically before it), degradation is not acting to reduce fluorescence. This offset is usually set to be the lowest molecular weight peak observed in one or more electropherograms (or some value below it). As the PCR occurs, some of the fluorescence that was destined for the allele will shift to stutter positions on the EPG. There are a number of stutter types that can occur (back stutter, forward stutter, half stutter, double stutter, etc.) and we will define the number of types of stutter as *I*, the stutter ratio of stutter type *i*, for locus *l* for allele *a* as $\pi^{l,i}_a$ and the position of

stutter type *i* relative to the parent peak, *a*, as Δ^i. We can now split the total allelic product into components of, respectively, allele and stutter by:

$$E^{l,\bar{i}}_{a,n,y} = \frac{T^l_{a,n,y}}{1+\sum_i \pi^{l,i}_a}$$

and

$$E^{l,i}_{a+\Delta^i,n,y} = \frac{T^l_{a,n,y}\pi^{l,i}_a}{1+\sum_i \pi^{l,i}_a}$$

where $\bar{i}$, indicates 'not *i*' and hence the allelic component. The total expected height of a peak at a locus, replicate and kit combination is then the sum of the stutter and allelic components of all individuals that fall on that allelic position:

$E^l_{a',y} = \sum_n E^{l,\bar{i}}_{a',n,y} + \sum_n \sum_i E^{l,i}_{a+\Delta^i,n,y}$ where *a* is chosen for each *i* so that $a+\Delta^i = a'$

Doing this for all contributors, alleles, loci, replicates and kits results in *Y* expected profiles, each of which has an observed counterpart, for which each peak height can be compared. Let **E** be the vector of expected peak heights. We assume independence of observed peak heights given the expected peak heights (shown in other work) so that $p(\mathbf{O}\,|\,\mathbf{S_j}, \boldsymbol{M}) = p(\mathbf{O}\,|\,\mathbf{E}) = \prod_y \prod_l \prod_a \Pr\left(O^l_{a,y}\,|\,E^l_{a,y}\right)$. Differences between observed and expected peak heights, $\Pr\left(O^l_{a,y}\,|\,E^l_{a,y}\right)$, are modelled by transforming the variable $O^l_{a,y}$ to $\log_{10}\left(\frac{O^l_{a,y}}{E^l_{a,y}}\right)$, where:

$$\log_{10}\left(\frac{O^l_{a,y}}{E^l_{a,y}}\right) \sim N\left[0, \left(1-\sum_i \mathbb{P}^i\right)\frac{\left(c^{\bar{i}}\right)^2}{E^l_{a,y}} + \sum_i \frac{\mathbb{P}^i\left(c^i\right)^2}{O^l_{a+\Delta^i,y}}\right]$$

Where $\mathbb{P}^i$ is the proportion of peak *a* that is stutter type *i*. The right-hand side of equation 5 signifies modelling using a normal distribution, in the form N[mean,variance]. Note that the c^2 parameters (for either allele or stutter) in the variance term have a prior gamma distribution modelled by:

$\left(c^i\right)^2 \sim \Gamma\left(\alpha^i, \beta^i\right)$

Prior distributions for c^2 terms and σ^2 are determined using a hierarchical Bayes model, run on a dataset of *C* single source DNA profiles. For these profiles a single genotype set (in this case of one contributor a genotype set can be considered a genotype, *G*) exists for each locus in each contributor and is provided to the analyses known information. The integral term from earlier can then be more simply expressed as $\int p(\mathbf{O}, \boldsymbol{M})d\boldsymbol{M}$, where the genotype set probability for each profile is a known constant value and omitted from the analysis.

We also include an additional parameter within **M** during the hierarchical Bayesian analysis, which a value of σ^2 for each profile (whereas in standard DNA profile analysis a fixed value for σ^2 is used).

Let $\boldsymbol{\alpha} = \left(\alpha^1, ..., \alpha^I\right)$, $\boldsymbol{\beta} = \left(\beta^1, ..., \beta^I\right)$ and $\mathbf{V}$ be the set of all variance hyper-parameters $\mathbf{V} = \left(\boldsymbol{\alpha}, \boldsymbol{\beta}, \lambda\right)$. We seek posterior distributions for elements of $\mathbf{V}$:

$p(\mathbf{V}, \mathbf{M} \mid \mathbf{O}) \propto p(\mathbf{O} \mid \mathbf{M}, \mathbf{V})\, p(\mathbf{M} \mid \mathbf{V})\, p(\mathbf{V})$ where:

$\left(c^i\right)^2_1, \left(c^i\right)^2_2, ..., \left(c^i\right)^2_C \mid \mathbf{O} \sim \Gamma\left(\alpha^i, \beta^i\right)$

The gamma distribution was chosen as it has properties that align with the expectations of the variance variables. It is bounded at the lower end by zero, at the upper end by infinity, and likely to be asymmetrical. We model:

$\sigma^2_1, \sigma^2_2, ..., \sigma^2_C \mid \mathbf{O} \sim Exp\left(\lambda\right)$

Th exponential distribution was chosen as it has properties that align with the expectations of the σ^2 variables. We desire imbalance within the profile between loci to be described by other parameters (namely template and degradation) preferentially to locus amplification efficiency. In a pristine DNA sample, loci should all amplify approximately equally well. In practise there will some locus amplification efficiency differences between loci due to extraneous chemicals carried through the DNA extraction process that affect amplification (which some loci are more sensitive to than others) and so some level of locus amplification efficiency variance is needed but should be minimised. However, if too much variance is seen then this is an indication of a poor DNA profile amplification and the profile should be considered unsuitable for analysis. As stated, in the standard analysis of DNA profile data we do not have a σ^2 parameter. Instead a constant value of $\sigma^2 = \lambda^{-1}$ is used. Within the hierarchical Bayesian analysis we trialled a gamma model instead of an exponential model for σ^2 terms but found little difference in outcome (data not shown) between the mean of the gamma distribution and the mean of the exponential distribution and so chose the simpler distribution.

The process for determining posterior distributions for variance terms is:

- For each of the *C* profiles, holding the values within **V** constant:
  - Draw values for parameters for profile *c* within $\mathbf{M_c}$ by random walk
  - Evaluate $p(\mathbf{O}_c \mid \mathbf{M}_c, \mathbf{V}) \Pr(\mathbf{M}_c, \mathbf{V})$
  - Accept or reject proposed parameters by Metropolis-Hasting algorithm
  - Repeat *X* times
- Holding all values within **M** constant:
  - Draw values for hyper-parameters within **V** by random walk
  - Evaluate $\Pr(\mathbf{V}) \prod_c p(\mathbf{O}_c \mid \mathbf{M}_c, \mathbf{V}) \Pr(\mathbf{M}_c \mid \mathbf{V})$
  - Accept or reject proposed parameters by Metropolis-Hasting algorithm
  - Repeat *X* times
- Repeat outer loops until converged

Values of X from 1 to 10000 were trialled but were found to make minimal difference to the resulting posterior distributions (data not shown). Prior distributions for parameters within **V** are:

$\alpha \sim U[1.5, \infty]$

$\beta \sim U[0, \infty]$

$\lambda \sim U[0, \infty]$

The restriction on the lower bound value for α comes from a desire to limit the shape of the gamma distribution to non-exponential curves. In practise, values for α are not obtained below this level even when the restriction is not placed on the prior.

Chapter 5: Testing the functioning of the models

Traditional methods of calculating a numerical value for the LR relied on models, simple enough to have closed set formula that could be calculated exactly. Typical 'validations' of software that implemented these models would then consist of calculating, by hand, a series of LRs for different scenarios and showing that the software provided the same answer. For the first time, with the introduction of STRmix™ forensic laboratories were faced with a stochastic system that:

- produced a different answer each time it was run
- used formulae that were not closed sets (for the most part being complex multiple integrals)
- produced numbers that couldn't be reproduced by hand and
- was designed to analyse DNA profiles that were traditionally considered beyond the ability of humans to interpret (prior to STRmix™ the general consensus amongst Australian government forensic DNA laboratories was that a mixture of 3 people of reasonable quality was the limit to interpretation. Now in 2017, with STRmix™ laboratories are analysing complex, low level, mixed DNA profiles originating from 5 people).

Questions arose from laboratories such as:

- How do I know if the system is giving the right answer?
- How can I check LRs that are too complex to replicate manually?
- How, as a human, do I assess a system designed to perform beyond my abilities of assessment?

These questions were echoed in defence questions in a number of court challenges to the STRmix™ methodology. The court questions tended to centre around themes of '*how do you know the system is reliable?*', or the blunter statement '*I put it to you that you are doing nothing but guessing*'.

Up until this point the main body of work that had been published had not focussed on validation, but rather defining, and testing models, and proposing extensions to LR calculations. When questioned on reliability, the main body of published work to which the forensic community could point was the validation section of the paper in section 2.6. A point of interest arose from this work, and in particular Figure 6 (from the paper in section 2.6), which showed the concordance between manual interpretations and the corresponding LR when STRmix™ was used. The forensic community often showed this graph as an example of how well the system was performing, with the vast majority of LRs being concordant with human interpretation and how much better the system was able to make use of data with LRs able to be provided for many more results than previously possible. The law community expressed concern with the graph, citing the fact that there were some (very few) instances where the continuous system gave contradicting results to the human interpretation and that it was able to 'make up' numbers even for profiles that, by existing forensic standards, were inconclusive. While the forensic community had moved to consider the continuous system the new gold standard and, seeing the graph, realised how far the forensic biology community had

come, the legal community still considered manual interpretations as the gold standard, and to them this graph simply showed that the new system was unreliable.

These driving forces prompted several different bodies of work. One body of work focussed on demonstrating how complex systems could be validated, not by scrutinising (or manually reproducing) individual results, but rather by considering the trends in the LRs that were expected over a range of problems with changing components. This work became the publication given in section 5.1.

Another body of work considered how other diagnostics could demonstrate the reliability of the LR. Theoretical expectations were derived for exceedance probabilities (the probability of choosing an individual from the population, who had not contributed DNA, and them yielding at least as much support for inclusion in the DNA mixture as the true donors) and average LR size, based on work by Allan Turing. This lead to the publication in section 5.2.

By the time the manuscript in section 5.2 was published, there were a number of continuous or semi-continuous DNA interpretation systems in existence (at present the count is approximately eight) that were grouped under the title 'probabilistic genotyping systems'. There was growing popularity in forensic DNA laboratories around the world for these systems and in response to the popularity, international advisory bodies started working on validation guidelines for probabilistic genotyping systems. These included the European based International Society of Forensic Genetics (ISFG) and the American based Scientific Working Group for DNA Analysis Methods (SWGDAM). The suggestion from some members of these groups was that that the Hd true tests (described in 5.2) should make up part of the recommended validation. Ultimately it was decided within the group not to make this a recommendation due to the impracticality of the size of the tests required. The argument went as follows:

Modern DNA profiling systems produce DNA profiles with frequencies less than 1 in 10^{20}. In order to properly test exceedance probabilities or average LRs from random draws of profiles from the population would therefore require $>10^{20}$ draws. No computer had the power or speed to complete such a task in a time that would be acceptable.

In response to this limitation, although it was too late for inclusion in the recommendations (which had already been published), work was carried out that demonstrated how importance sampling could be applied to bias the choice of profile chosen and then adjust afterwards to recover diagnostics of interest. This work led to the publication in section 5.3. Importance sampling is now a feature available in STRmix™, if users wish to carry out such testing against a deconvolution.

Around the same time, the opinion was voiced that when DNA profiles became weak or peak heights became highly variable (as is the case in certain DNA profiling workflows), peak heights presented no further information beyond the presence or absence of the peaks

themselves. There was a relatively simple way in which this assertion could be tested. STRmix™ was provided with the same set of very low level and complex DNA profiles twice, once using its full capabilities and once in a modified version of the program that did not use peak heights (and working in a way similar to a semi-continuous model). The ability of both analyses to provide support for the contribution of the known donors of DNA was then compared. This work drove the publication given at the end of this chapter that showed there was still information in low-level peaks that could be utilised by a fully continuous DNA profile interpretation system.

Manuscript: Using continuous DNA interpretation methods to revisit likelihood ratio behaviour. D Taylor. (2014) Forensic Science International: Genetics 11, 144-153 – *Cited 16 times*

Statement of novelty: This work revisited some old theories on LR behaviour, extended by the application of new, continuous DNA interpretation methods to demonstrate their relevance in a contemporary setting.

My contribution: Sole author

Research Design / Data Collection / Writing and Editing = 100% / 100% / 100%

Additional comments:

Forensic Science International: Genetics 11 (2014) 144–153

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

Using continuous DNA interpretation methods to revisit likelihood ratio behaviour

Duncan Taylor *

Forensic Science South Australia, 21 Divett Place, Adelaide, SA 5000, Australia
School of Biological Sciences, Flinders University, GPO Box 2100 Adelaide SA 5001, Australia

ARTICLE INFO

Article history:
Received 22 January 2014
Received in revised form 4 March 2014
Accepted 13 March 2014

Keywords:
DNA profile interpretation
Mixtures
Likelihood ratio
Assumed contributors
STRmix
Reliability

ABSTRACT

Continuous DNA interpretation systems make use of more information from DNA profiles than analysts have previously been able to with binary, threshold based systems. With these new continuous DNA interpretation systems and a new, more powerful, DNA profiling kit (GlobalFiler) there is an opportunity to re-examine the behaviour of a commonly used statistic in forensic science, the likelihood ratio (*LR*). The theoretical behaviour of the *LR* has been known for some time, although in many instances the behaviour has not been able to be thoroughly demonstrated due to limitations of the biological and mathematical models being used. In this paper the effects of profile complexity, replicate amplifications, assuming contributors, adding incorrect information, and adding irrelevant information to the calculation of the *LR* are explored. The empirical results are compared to theoretical expectations and explained. The work finishes with the results being used to dispel common misconceptions around reliability, accuracy, informativeness and reproducibility.

1. Introduction

Situations arise commonly in forensic DNA casework where an analyst will need to make a choice whether a profile should be interpreted. If so, further choices regarding appropriate assumptions and propositions will be needed. Ultimately these choices will lead to the calculation of a statistical weighting, commonly a likelihood ratio (*LR*), which answers some question of interest for the analyst, police or court.

In order to make these choices, the analyst must use their knowledge of the expected behaviour of the *LR* under different circumstances. This would include the way it reacts to:

- Increasing DNA profile complexity, either through number of contributors or profile quality, and
- Assumptions of contributors known or assumed to be present under both propositions, and
- Replicate amplifications.

The theoretical behaviour of the *LR* has been known for some time [1], although in many instances the behaviour has not been able to be thoroughly demonstrated practically due to limitations of the biological and mathematical models being used. Continuous models utilise more information from DNA profiles than previous, binary, models and so are able to demonstrate these *LR* behaviours.

The continuous system of Taylor [2] was used to demonstrate some well understood behaviours of the *LR*, such as the effect of increasing profile complexity, adding replicate amplifications and assuming contributors to a profile. Also explored were behaviours which have not previously been demonstrated practically. These included the effect of adding incorrect information into the calculation (i.e. when both propositions of the *LR* are false), and the effect of adding irrelevant information to the calculation.

The results obtained provided a useful review of the behaviours of the *LR* using empirically obtained data rather than mathematical theory. It is hoped that the examples given in this work will inform analysts required to make casework decisions.

2. Methods

2.1. Experimental setup

DNA was obtained from four individuals with informed consent and used to construct all mixtures. PCR amplifications were carried out using GlobalFiler (Life Technologies) as per manufacturer's instructions.

Allele frequencies were derived from an in-house database comprising 186 self-declared South Australian Caucasian individuals (database validation not shown).

* Tel.: +61 8 8226 7700; fax: +61 8 8226 7777.
E-mail address: Duncan.Taylor@sa.gov.au

http://dx.doi.org/10.1016/j.fsigen.2014.03.008
1872-4973/

Table 1
Mixture proportions and PCR setup.

Tubes	Mixture proportions for contributor				Total DNA added to PCR (pg)
	One	Two	Three	Four	
1–3	0.50	0.50			400,200,50
4–6	0.33	0.67			
7–9	0.20	0.80			
10–11	0.17	0.83			
13–15	0.09	0.91			
16–18	0.33	0.33	0.33		
19–21	0.50	0.33	0.17		
22–26	0.25	0.25	0.25	0.25	400,200,50,20,10
27–31	0.40	0.30	0.20	0.10	

The 186 individuals used for allele frequency generation and the four individuals used to create the mixtures were compiled into a 190 individual database for comparison to all constructed mixtures.

Two, three and four person mixtures were constructed in varying proportions and amplified with varying amounts of template DNA as described in Table 1. Mixtures were prepared from stock solution of each contributor, made up to known concentration by quantification on an ABI PRISM® 7500 Sequence Detection System using Quantifiler™ Human DNA Quantification Kit (Life Technologies). Stock solutions were quantified twice and an average taken for the final value. Each experimental setup was amplified in triplicate giving a total of 93 profiles.

Profiles were analysed using software STRmix which utilises models described in [2–4] (exact software settings used are available from the author on request). In all analyses the Y-indel locus and DYS391 were ignored.

2.2. LR *calculations and figures*

LR calculations were performed using propositions as described in each experiment below. For all calculations the product rule was used (i.e. no co-ancestry coefficient) and the point estimate has been given. This is not the recommended practice for criminal casework but allows the other effects to express themselves.

LR calculations in the experiments listed below considered each person on the 190 individual database as a potential contributor, or person of interest (POI), to the mixed DNA profiles. In doing so there are comparisons to all individuals who are known to have contributed (when H_1 is true) to the DNA profile and the remainder, who are known not to have contributed (when H_2 is true).

Figs. 1–9 show the $\log_{10}(LR)$ produced for these comparisons. The *LR*s produced from comparisons to known contributors are signified by a blue point and those produced from comparisons to known non-contributors are signified by a red point. A minimum value for $\log_{10}(LR)$ of −30 was used, and any *LR*s obtained that fell below this were given the value of −30. The lines on figures were produced from LOWESS [5,6] and are given only as a visual indication of trends in the scattered results.

The polygons seen in some figures give a visual indication of the spread of *LR*s. They are produced by connecting the maximum and minimum values within each DNA amount bracket (as indicated on the *x*-axis of figures) at the midpoint of that bracket (and hence some points fall just outside the drawn polygon).

The amount of DNA contributed by each known contributor was known from the experimental design. When comparing to non-contributors, the choice of input DNA (for Figs. 1–9) was not known as the non-contributor could align with any of the contributors' input DNA amounts. For known non-contributors the amount of input DNA was assigned as the total amount of DNA added to the PCR divided by the number of contributors. This was likely to be an overestimation of input DNA amount for the non-contributors as many of the *LR* values obtained would be when the individuals are aligned with the smaller contributor in the profile. Hence the red points on Figs. 1–9 have fewer DNA amount values than the blue points. For example tube 8 in Table 1 has total DNA amount of 200 pg and mixture proportions 0.2 and 0.8. When an *LR* was calculated for the comparison to contributor 1 it was placed at 40 pg and for contributor 2 the *LR* was placed at 160 pg. For all comparison to unknowns the *LR*s were placed at 100 pg, being the average input DNA amount of contributors to the profile.

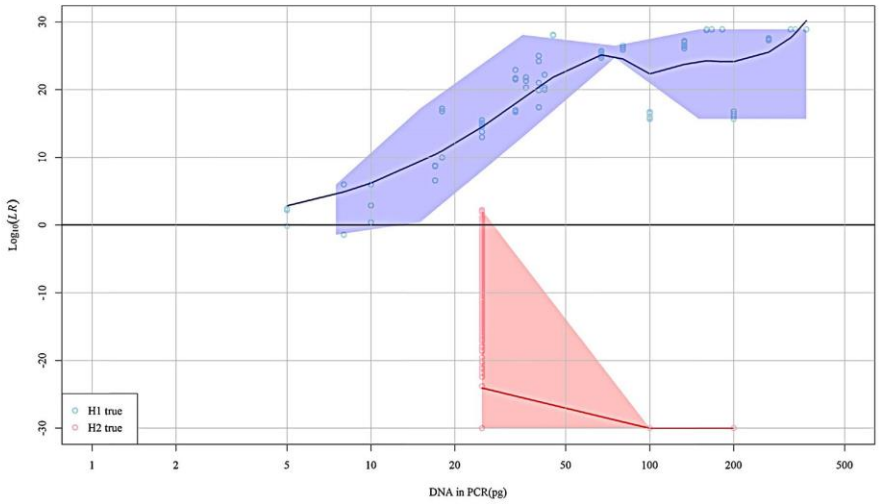


Fig. 1. Experiment 1 – LRs produced for two person mixtures, with LOWESS lines and polygons showing coverage of scatterplot points. (For interpretation of the references to color in the text, the reader is referred to the web version of the article.)

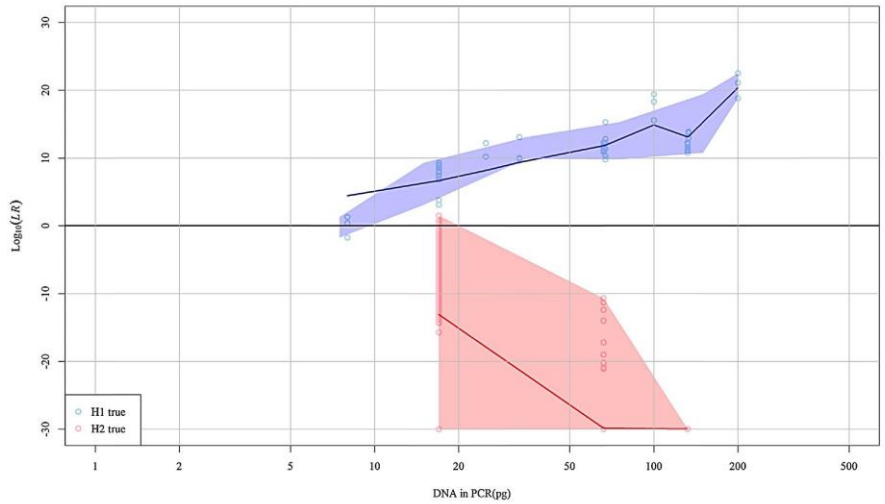


Fig. 2. Experiment 1 – LRs produced for three person mixtures, with LOWESS lines and polygons showing coverage of scatterplot points. (For interpretation of the references to color in the text, the reader is referred to the web version of the article.)

2.3. Experiment 1: standard deconvolutions

In experiment 1 all 93 profiles were individually analysed using their correct number of contributors. *LR*s were calculated using propositions for an *N* person mixture:

H_1. The POI and $(N-1)$ unknown individuals are the sources of DNA

H_2. *N* unknown individuals are the sources of DNA

2.4. Experiment 2: the power of multiple PCRs

In experiment 2 the 10 sets of four person mixtures were analysed using all three replicate PCRs in each analysis (as opposed to Experiment 1 where they were analysed separately). The propositions used were:

H_1. The POI and 3 unknown individuals are the sources of DNA

H_2. 4 unknown individuals are the sources of DNA

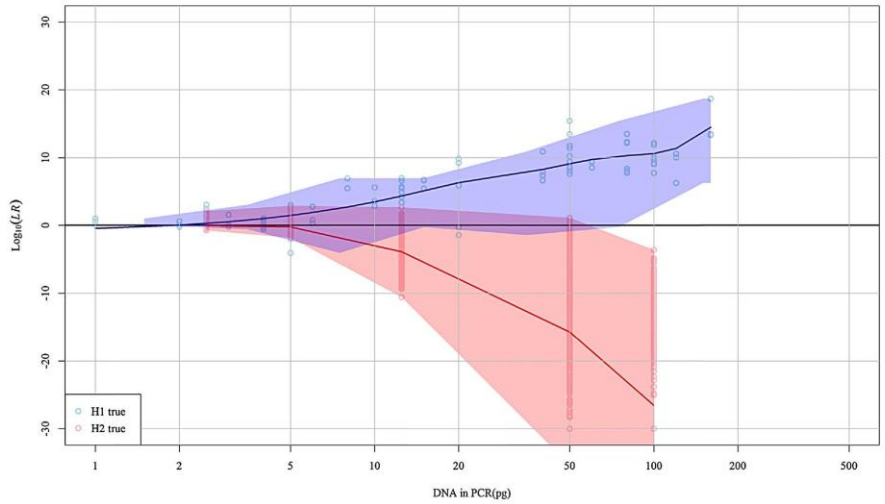


Fig. 3. Experiment 1 – LRs produced for four person mixtures, with LOWESS lines and polygons showing coverage of scatterplot points. (For interpretation of the references to color in the text, the reader is referred to the web version of the article.)

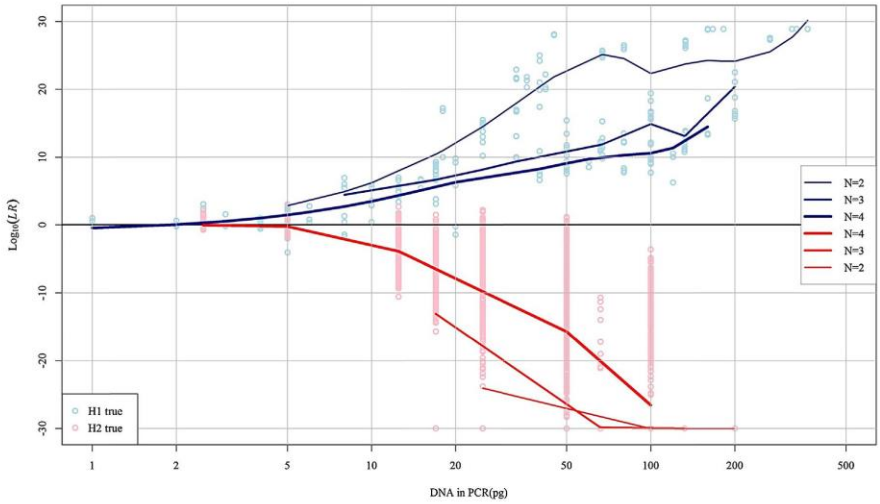


Fig. 4. LRs and LOWESS lines from Fig. 1 ($N = 2$), 2 ($N = 3$) and 3 ($N = 4$) overlayed. (For interpretation of the references to color in the text, the reader is referred to the web version of the article.)

2.5. Experiment 3: adding correct information

In experiment 3 the 10 sets of three PCR four person mixtures were analysed, this time assuming three out of the four known contributors. Each combination of three individuals was assumed, meaning from the 10 sets of PCRs, 40 analyses were carried out and compared to POI using propositions:

H_1. The POI, contributor A, contributor B and contributor C are the sources of DNA

H_2. Contributor A, contributor B and contributor C and an unknown individual are the sources of DNA

Where A, B and C were combinations of contributors 1, 2, 3 or 4 omitting one at a time. *LR*s were not produced for contributors who were assumed to be present.

2.6. Experiment 4: adding incorrect information

In experiment 4 the 10 sets of three PCR, four person mixtures were analysed, this time assuming an artificially constructed non-contributor. The 'fake' references (fakeREF) were constructed so that they might not be excluded from the mixtures by typical human interpretation method. This was achieved by constructing the fakeREFs using genotypes randomly chosen from the four

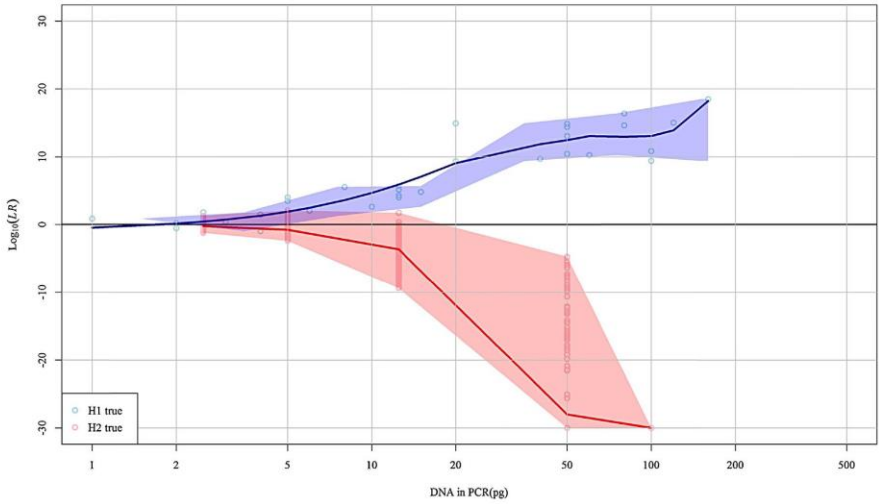


Fig. 5. Experiment 2 – LRs produced for four person mixtures using three amplifications, with LOWESS lines and polygons showing coverage of scatterplot points. (For interpretation of the references to color in the text, the reader is referred to the web version of the article.)

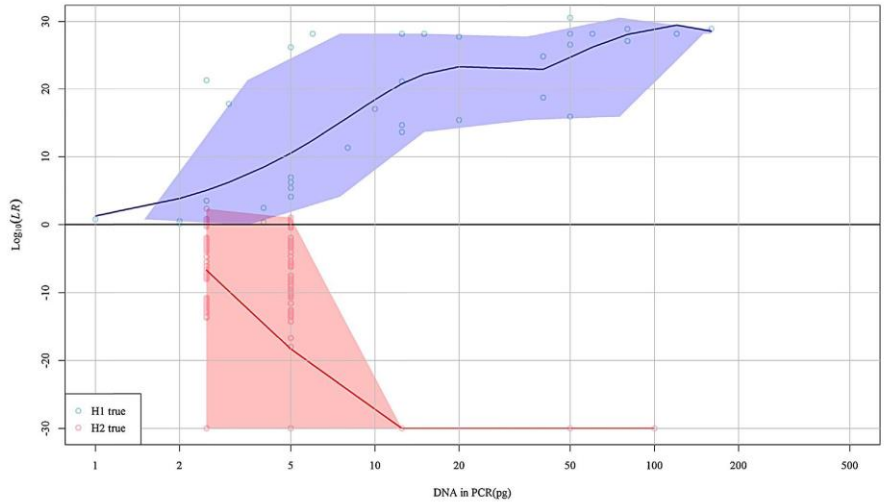


Fig. 6. Experiment 3 – LRs produced for four person mixtures using three amplifications and assuming three out of the four known contributors in each analysis, with LOWESS lines and polygons showing coverage of scatterplot points. (For interpretation of the references to color in the text, the reader is referred to the web version of the article.)

known contributors for each locus. A different fakeREF was generated for each of the 10 analyses, assumed to be a contributor and compared to POIs using propositions:

H_1. The fakeREF, POI and 2 unknown individuals are the sources of DNA

H_2. The fakeREF and 3 unknown individuals are the sources of DNA

2.7. Experiment 5: adding irrelevant information

In experiment 5 the 15 sets of three PCR, two person mixtures were analysed as three person mixtures but assuming a randomly chosen non-contributor from the searchable database (dbREF) was a contributor. Doing this had the effect of adding an additional contributor but effectively forcing their contribution to the profile to be close to zero, hence the additional information was

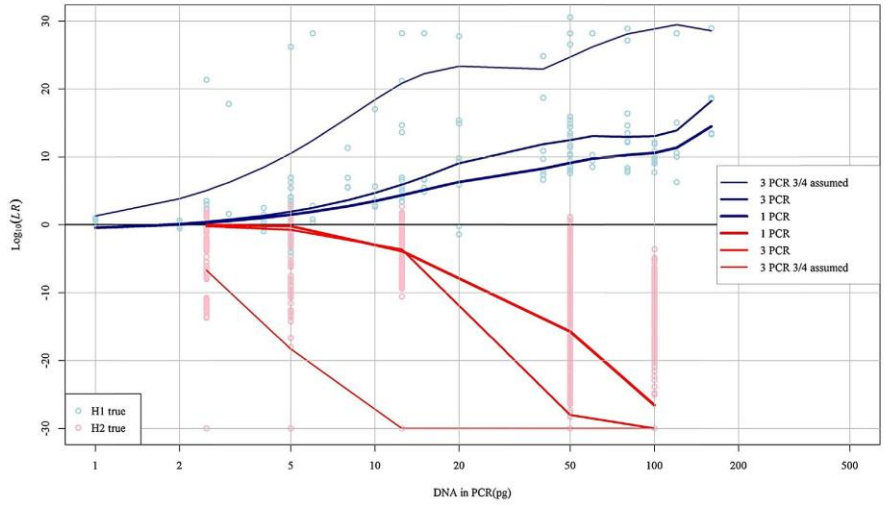


Fig. 7. LRs and LOWESS lines from Fig. 3 (1 PCR), 5 (3 PCR) and 6 (3 PCR 3/4 assumed) overlayed. (For interpretation of the references to color in the text, the reader is referred to the web version of the article.)

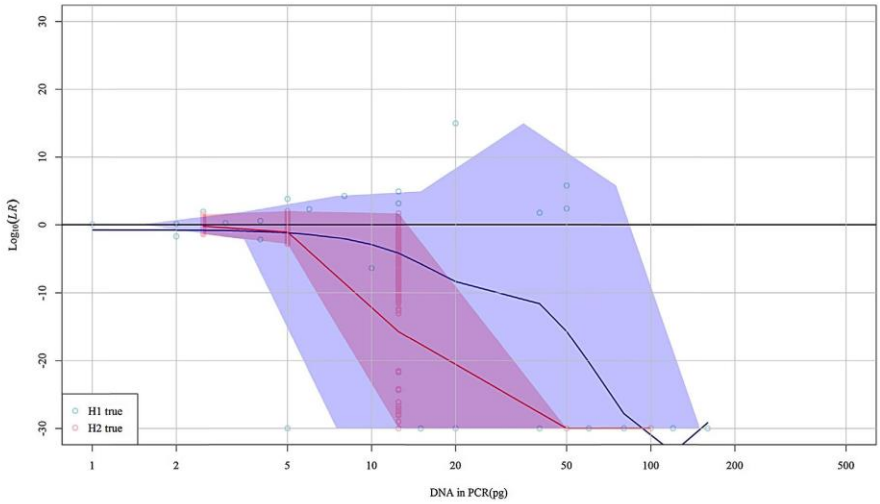


Fig. 8. Experiment 4 – LRs produced for four person mixtures using three amplifications and assuming an artificially constructed non-contributor in each analysis, with LOWESS lines and polygons showing coverage of scatterplot points. (For interpretation of the references to color in the text, the reader is referred to the web version of the article.)

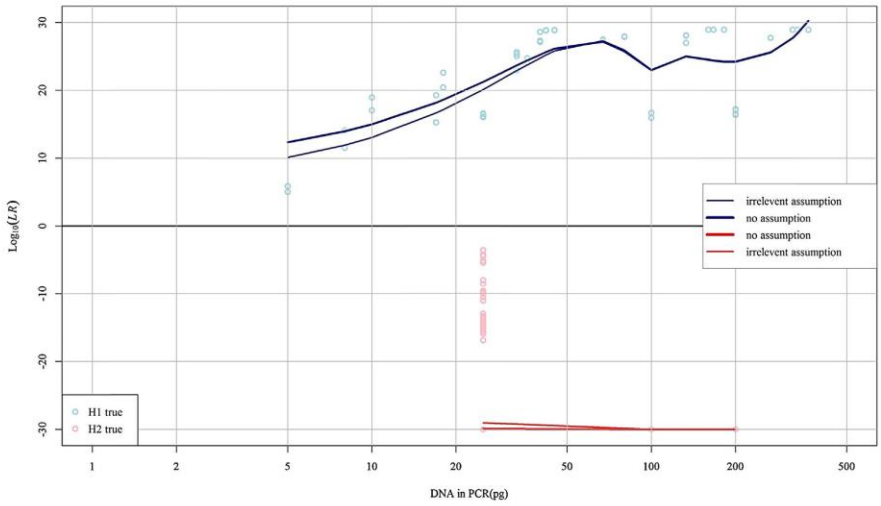


Fig. 9. Experiment 5 – LRs produced for two person mixtures using three amplifications (no assumption) and analysed as three person mixtures and assuming a non-contributor (irrelevant assumption), with LOWESS lines. (For interpretation of the references to color in the text, the reader is referred to the web version of the article.)

effectively irrelevant to the original calculations. *LR*s were calculated using propositions:

H_1. The dbREF, POI and an unknown individuals are the sources of DNA

H_2. The dbREF and 2 unknown individuals are the sources of DNA

The *LR*s produced were compared to those produced by analysing the 15 sets of three PCR two person mixtures analysed using the propositions:

H_1. The POI and an unknown individuals are the sources of DNA

H_2. 2 unknown individuals are the sources of DNA

3. Results and discussion

Before examining the empirical results it is worth briefly re-examining the theory behind *LR* calculations, so that the results can be compared to the expected behaviour.

3.1. LR theory

The *LR* considers the probability of obtaining some evidence given two competing propositions, H_1 and H_2. When applied to DNA profile evidence the *LR* is the ratio of two sums:

$$LR = \frac{\sum_j w_j Pr(S_j|H_1)}{\sum_{j'} w_{j'} Pr(S_{j'}|H_2)}$$

where, using the nomenclature of Taylor [2], w_j is a weight for the jth genotype set, S_j, being considered under the propositions. The sums can be across a single element, if the genotype of the contributors can be assigned unambiguously, or numerous elements, and in a continuous system the weights can take any non-negative value (note that in Taylor [2] weights are values between 0 and 1 as they are normalised. Weights will be referred to in the normalised form for the remainder of this work). This *LR* construct can be used to consider how information utilised in an *LR* calculation will affect its magnitude.

If the weights that correspond to genotypes sets considered in H_1 are increased, relative to all possible genotype sets (commonly in forensic calculations, those considered in H_2) then the *LR* will increase. These weights are increased when there is more certainty about their corresponding genotype sets being the source of the observed profile.

Consider the single locus, two person profile in Fig. 10, with known sources (13,14) and (16,17). There are six possible genotype sets that could explain the observed profile as shown in Table 2.

Depending on peak heights, both absolute and relative to each other, some of the genotype sets will be poor descriptions of the observed profile. Normally DNA profiles being analysed are multi-locus and the other loci will dictate contributor positions for major or minor components to the profile (if they exist). Assume that contributor position 1 is the major component and position 2 is the minor component. If the peak heights in Fig. 10 are sufficiently intense enough that the imbalances between the 13 or 14 and the 16 or 17 are very unlikely then the weights would be $w_1 = 1$ and all others equal to zero. In this situation all the genotypic probability is distributed to the genotype set that corresponds to the known contributors. If the known source (referred to in the propositions as the POI) (13,14) was to be compared to this profile using propositions:

H_1. The POI and 1 unknown individual are the sources of DNA

H_2. 2 unknown individuals are the sources of DNA

Then the *LR* would be:

$$LR = \frac{w_1 \times 2p_{16}p_{17}}{w_1 \times 2p_{13}p_{14} \times 2p_{16}p_{17}} = \frac{1}{2p_{13}p_{14}}$$

Now consider adding uncertainty into the scenario seen in Fig. 10 by imagining that the peak heights were very close together and low in intensity. Doing so results in a situation where $w_1 = w_2 = \ldots = w_6 = w$, so that the genotypic probability has been distributed amongst six different genotype sets and hence away from the genotype set of the true contributors compared to the previous scenario. Using the same propositions as before the *LR* would now be:

$$LR = \frac{w \times 2p_{16}p_{17}}{w \times 2p_{13}p_{14} \times 2p_{16}p_{17} + \ldots + w \times 2p_{14}p_{16} \times 2p_{13}p_{17}} = \frac{w \times 2p_{16}p_{17}}{6w \times 4p_{13}p_{14}p_{16}p_{17}} = \frac{1}{6 \times 2p_{13}p_{14}}$$

Being less certain about the genotypes of the contributors has distributed the genotype probability away from the genotype set corresponding to the known contributors and over a number of other genotype sets. This has had the result that the *LR* has dropped by a factor of 6. This is the Evett et al. [7] result.

Now consider the scenario as above with low level, equal peak heights but assume that the known contributor 2, with genotype (16,17), is present. Adding this information has the effect of distributing the genotypic probability all back to genotype set 1 so that again $w_1 = 1$ and all others equal to zero. The *LR* in this instance will again be:

$$LR = \frac{1}{2p_{13}p_{14}}$$

This is a demonstration that removing information from the analysis decreases the expected *LR* for a true proposition and adding correct information increases the expected *LR* for a true proposition.

Assume now instead a non-contributor to the profile seen in Fig. 10, an individual who is (13,16). In this instance all the weight is on a single genotype set, but this time $w_2 = 1$, which means the weight for the genotype set corresponding to the true contributor, $w_1 = 0$. The *LR* for comparison to known contributor 1 (13,14) would be:

$$LR = \frac{0}{2p_{14}p_{17}} = 0$$

And a similar result would be obtained for the comparison to known contributor 2 (16,17).

Providing incorrect information to the *LR*, by assuming a non-contributor is present, has led to the exclusion of the true contributors.

3.2. Summary of findings

Continuous systems of DNA interpretation have allowed the behaviour of *LR* calculations to be demonstrated using empirical

Table 2
Genotypes sets for profile seen in Fig. 10.

Set	Weight	Contributor genotype	
		Position 1	Position 2
1	w_1	(13,14)	(16,17)
2	w_2	(13,16)	(14,17)
3	w_3	(13,17)	(14,16)
4	w_4	(16,17)	(13,14)
5	w_5	(14,17)	(13,16)
6	w_6	(14,16)	(13,17)

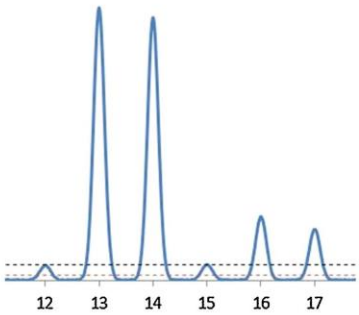


Fig. 10. Example EPG showing a single locus of a two person mixture.

data. Many of these behaviours are known in a theoretical sense, however, the demonstration of them visually, and collected together in one paper provides a useful reference for DNA analysts. This work also demonstrates the magnitude of the effects that different assumptions and propositions can have on the calculation of an *LR*, something that would be difficult to portray from theory alone. Also, although it is not the focus of this paper, the work conducted demonstrates the power of continuous DNA interpretation systems to analyse DNA profile data.

3.3. Experiment 1

As DNA profiles are generated from less DNA the *LR* produced comparing known contributors trends down towards one and the *LR* produced comparing known non-contributors trends upwards towards one. This demonstrates that the ability for the *LR* to distinguish between a true and a false proposition is reduced as less correct and relevant information is provided to the calculation. Less information can either be fewer peaks or lower intensity peaks. Lower peak intensity is a reduction in information as there is less certainty that the peak height is representative of the input DNA amounts. These effects are demonstrated in all figures.

As DNA profiles become more complex, by increasing the number of contributors from which they are generated, the *LR*s trend towards one. Figs. 1–3 show *LR*s generated when comparing individuals to two, three and four person mixtures respectively. This is a further example of the point above, demonstrating a reduction in information, this time per contributor. The information reduction per contributor arises because as the number of contributors increases, so too does the number of genotype sets that can explain the observed data. The genotype set probability is therefore distributed amongst more genotype sets in complex profiles. Fig. 4 shows the effect of additional contributors on the *LR*, reducing its ability to distinguish between true and false propositions as *N* increases.

3.4. Experiment 2

As more replicate biological analyses are concurrently used the *LR* generated for comparison to known contributors trends away from one upwards and the *LR* generated for comparison to known non-contributors trends away from one downwards. This is a further example of providing more correct and relevant information to the analyses. The effect is to have more certainty in the peaks height's representation of input DNA amount and hence distribute more probability to the weights associated with the true contributor's genotypes. Fig. 5 shows the effect of multiple PCRs, particularly when compared to Fig. 3, which is the same data analysed using individual replicates per analysis.

3.5. Experiment 3

As more known contributors are assumed to be present in a mixed DNA profile, the *LR* generated for comparison to known contributors trends away from one upwards and the *LR* generated for comparison to known non-contributors trends away from one downwards. Again this is an example of providing more correct and relevant information to the analyses. Assuming true contributors removes many of the otherwise possible genotype sets, leaving a very restricted list of genotype sets that includes the genotype set of the known contributor(s). Therefore the weights are concentrated on true contributor's genotype. Fig. 6 shows the results of assuming correct information. Fig. 7 shows the effects of providing incrementally more correct information to the *LR*, by considering four person profiles individually, in triplicate and in triplicate with the correct assumption of known contributors.

Also note that the ability of the *LR* to distinguish between true and false propositions is improved more by the assumption of known contributors than by the addition of replicate PCRs. The reason for this can be explained by considering a profile originating from equally contributing individuals. No number of replicate PCRs is going to provide additional resolution to the genotypes of contributors. However, providing one or more of the true contributor's profiles can dramatically reduce the number of genotype sets able to sensibly describe the EPG(s), and hence increase resolution of the remaining contributor(s) genotypes. This effect was observed in Fig. 7.

3.6. Experiment 4

If a known non-contributor is assumed to be present in a mixture then the effect on the *LR* can range from very little, to complete exclusion of known contributors. This is an example of providing incorrect information to analyses. The results in Fig. 8 show the dramatic effect that providing incorrect information to an *LR* can have, particularly for comparisons to known contributors. The reason for this dramatic and varied effect on the *LR* is that the assumption of a non-contributor will reduce the number of genotype sets the genotype probability is distributed across, but in doing so it may force allelic pairings that distributes the probability away from the genotype set that corresponds to the true contributors.

3.7. Experiment 5

If a non-contributor is assumed to be present in an n person mixture, and the mixture is analysed as originating from $n+1$ individuals then there is very little effect on the *LR* when comparing known contributors. The addition of the contributor allows the analyses to assign a near zero contribution for the assumed individual and remaining contributor genotypes are treated as though the assumed contributor was not present. Table 3 shows the mixture proportions assigned by the software to each contributor during the analyses, compared with the known mixture proportions. In the third set of proportions it can be seen that, as expected, the proportion assigned to the wrongly assumed contributor is very low.

This is an example of providing irrelevant information to analyses. Fig. 9 shows the effect on the *LR* when irrelevant information is provided. In Fig. 9 it can be seen that the effect on the *LR* when adding irrelevant information is negligible at the higher end of DNA contribution. At lower DNA contribution levels there is a divergence between the two sets of results, with the irrelevant assumption LOWESS line falling below the no assumption LOWESS line. At these lower levels of input DNA the small mixture proportion assigned to the assumed contributor is in the same region as the known minor contributor (see tubes 13, 14 and 15 in Table 3). The effect is for the weights to be spread amongst a larger number of genotypes sets as the assumed contributor can account for some of minor peaks that they share, by chance, with the known minor contributor. At these lower levels therefore experiment 5 is no longer demonstrating irrelevant information as it is having an effect, albeit small, on the weights.

3.7.1. The difference between reliable, reproducible and informative

Having presented and explained DNA interpretation to scientists, lawyers and juries it has been the author's experience that the terms reliable, reproducible and informative often get used incorrectly, and this can lead to the wrong conclusions being drawn. The excerpt below is a transcript of a defence expert's testimony who was challenging some *LR* results generated from

Table 3
Mixture proportions (Mx) obtained in Experiment 5.

Tube	DNA in PCR (pg)	Designed Mx for contributor		Mx obtained when analysed as $N=2$		Mx obtained when analysed as $N=3$		
		One	Two	One	Two	One	Two	Three (assumed)
1	400	0.50	0.50	0.50	0.50	0.49	0.49	0.02
2	200	0.50	0.50	0.50	0.50	0.49	0.49	0.02
3	50	0.50	0.50	0.50	0.50	0.46	0.47	0.07
4	400	0.33	0.67	0.32	0.68	0.31	0.65	0.04
5	200	0.33	0.67	0.31	0.69	0.30	0.67	0.03
6	50	0.33	0.67	0.41	0.59	0.48	0.49	0.03
7	400	0.20	0.80	0.17	0.83	0.16	0.82	0.02
8	200	0.20	0.80	0.20	0.80	0.19	0.78	0.03
9	50	0.20	0.80	0.21	0.79	0.18	0.76	0.06
10	400	0.17	0.83	0.16	0.84	0.15	0.83	0.02
11	200	0.17	0.83	0.19	0.84	0.16	0.82	0.02
12	50	0.17	0.83	0.08	0.92	0.12	0.82	0.06
13	400	0.09	0.91	0.06	0.94	0.07	0.90	0.03
14	200	0.09	0.91	0.08	0.92	0.07	0.09	0.03
15	50	0.09	0.91	0.04	0.96	0.04	0.90	0.06

low level data using the method in Taylor 2013 [2]. The specific use of reliability referred to the lab hardware in this example:

> ...One cell has got 7 picograms of DNA inside it, so we are looking at the results of about four cells worth of DNA. Now, that is beyond the reliable capability of the (ed) amplification kits.

Another example here shows a question from defence council who has used the word 'accuracy' when 'uninformative' would be better. He was referring to a graph that showed peak height variability increasing as peak heights decrease:

> Defence: The conclusion to draw is that as you get close to lower peak height, then the accuracy of the STRmix program drops?
>
> Witness: No, it's not what I would take from that. Within STRmix – I'll step back. This sort of modelling shows that as peak heights decrease, the variability increases, so if you were to generate multiple profiles from the same DNA extract and they contain low peak heights, the peak heights are going to be quite variable at low levels and not so variable at high levels. So we put models into STRmix. We tell it when you see low level peaks they could be quite variable, when you see high level peaks they're not going to be quite so variable. As STRmix is analysing the profiles, it accounts for a lot of possible variability, so it's going to give less strength to its predictions based on increased variability. That's the advantage of having these sort of models sitting behind STRmix.

The results from the work in this paper provide an opportunity to discuss the differences between these terms and relate them to practical results.

Reliable DNA results are those that can be trusted or used in an interpretation, statistical analysis or calculation. The reliability of DNA profiling results will depend on the quality assurance of a laboratory and the processes used to track, examine and biologically analyse an exhibit. The reliability of the statistical analysis or calculation will depend on the biological and mathematical models used.

Reproducibility is often used as part of dictionary definitions of reliability. This causes problems for forensic scientists because DNA profiles are not reproducible. Numerous works have studied the causes [8] and results [3,9–11] of the differences between EPGs produced from 'identical' duplicate analyses of the same DNA sample. Continuous DNA interpretation methods [2,12] also have a level of non-reproducibility as Markov Chain Monte Carlo systems are based on random number generation and so the statistic calculated differs each time they are run. The variation between replicate biological analyses does not, however, make the results unreliable as long as that variability is taken into account within the biological and mathematical models used to interpret them. Examples of statistical models that take high levels of non-reproducibility into account are the construction of consensus profiles for low template DNA analyses [13,14]. The reverse is also true; a completely reproducible result may be unreliable if the means in which it was generated are unreliable in some way.

Whether a result is informative relates to the outcome of the calculation and is not dependent on reproducibility or reliability (although hopefully any result obtained is done so reliably). In Fig. 3 at a high DNA input amount, the *LR*s produced when H_1 was true were separated from those produced when H_2 was true by many orders of magnitude. As the input DNA amount was reduced the *LR*s for both sets of data contracted around one, or $log_{10}(LR) = 0$. This is the expected result, because as the calculation has less information its power to distinguish between true and false propositions is reduced. Therefore, at the lower end of input DNA the results are considered uninformative, in that they do not provide the analyst a result that informs them that one proposition is supported over the other. Note that this does not mean that the result is unreliable or inaccurate, in fact the opposite is true, it is very reliably and accurately giving an uninformative *LR* of 1.

A second point that is commonly used, incorrectly, as an example of unreliability is that at lower levels of input DNA there are instances of *LR*s that falsely favour exclusion (blue points below the zero line in figures) and *LR*s that falsely favour inclusion of non-contributors (red points above the zero line in figures), commonly called adventitious matches. Adventitious matching is a phenomenon that has been known for some time [15,16] and recently work has been done using the method of Taylor 2013 [2] to investigate the rates of adventitious matching for complex and low level mixtures [17]. In fact the *LR* calculation is based on the premise that unrelated people can exist that possess the same DNA profile, purely by chance. So if an infinite number of non-contributors could be compiled in a database and compared to the analyses of mixtures in this paper, then the red points would range up to the same level as the blue points (and in some instances above them). Hence an overlap of *LR*s when H_1 is true and H_2 is true is not only an expected result, and fundamentally required for the *LR* to function, but is definitely not an indication of unreliability. It is the magnitude of the *LR* that gives an indication of the likely chance of adventitious matching having occurred.

Acknowledgements

I would like to thank Ian Evett and David Balding for their helpful discussions on this topic. I would also like to thank Chris Hefford for his work preparing the profile data. Points of view in this document are those of the author and do not necessarily represent the official position or policies of Forensic Science SA.

References

[1] D.V. Lindley, Understanding Uncertainty, John Wiley & Sons Inc., Hoboken, New Jersey, 2006.
[2] D. Taylor, J.-A. Bright, J. Buckleton, The interpretation of single source and mixed DNA profiles, Forensic Sci. Int. Genet. 7 (2013) 516–528.
[3] J.-A. Bright, D. Taylor, J.M. Curran, J.S. Buckleton, Developing allelic and stutter peak height models for a continuous method of DNA interpretation, Forensic Sci. Int. Genet. 7 (2013) 296–304.
[4] J.-A. Bright, D. Taylor, J.M. Curran, J.S. Buckleton, Degradation of forensic DNA profiles, Aust. J. Forensic Sci. 45 (4) (2013) 445–449.

[5] W.S. Cleveland, Robust locally weighted regression and smoothing scatterplots, J. Am. Stat. Assoc. 74 (1979) 829–836.
[6] W.S. Cleveland, LOWESS: a program for smoothing scatterplots by robust locally weighted regression, Am. Stat. 35 (1981) 54.
[7] I.W. Evett, C. Buffery, G. Willott, D.A. Stoney, A guide to interpreting single locus profiles of DNA mixtures in forensic cases, J. Forensic Sci. Soc. 31 (1991) 41–47.
[8] P. Gill, J. Curran, K. Elliot, A graphical simulation model of the entire DNA process associated with the analysis of short tandem repeat loci, Nucleic Acids Res. 33 (2005) 632–643.
[9] J.-A. Bright, K. McManus, S. Harbison, P. Gill, J. Buckleton, A comparison of stochastic variation in mixed and unmixed casework and synthetic samples, Forensic Sci. Int. Genet. 6 (2012) 180–184.
[10] J.-A. Bright, E. Huizing, L. Melia, J. Buckleton, Determination of the variables affecting mixed MiniFiler(TM) DNA profiles, Forensic Sci. Int. Genet. 5 (2011) 381–385.
[11] J.-A. Bright, J. Turkington, J. Buckleton, Examination of the variability in mixed DNA profile parameters for the Identifiler(TM) multiplex, Forensic Sci. Int. Genet. 4 (2009) 111–114.
[12] M.W. Perlin, M.M. Legler, C.E. Spencer, J.L. Smith, W.P. Allan, J.L. Belrose, et al., Validating TrueAllele® DNA mixture interpretation, J. Forensic Sci. 56 (2011) 1430–1447.
[13] P. Gill, J.P. Whitaker, C. Flaxman, N. Brown, J.S. Buckleton, An investigation of the rigor of interpretation rules for STR's derived from less that 100 pg of DNA, Forensic Sci. Int. 112 (2000) 17–40.
[14] J.-A. Bright, P. Gill, J. Buckleton, Composite profiles in DNA analysis, Forensic Sci. Int. Genet. 6 (2012) 317–321.
[15] B.S. Weir, Matching and partially-matching DNA profiles, J. Forensic Sci. 49 (2004) 1009–1014.
[16] T. Tvedebrink, P.S. Eriksen, J.M. Curran, H.S. Mogensen, N. Morling, Analysis of matches and partial-matches in a Danish STR data set, Forensic Sci. Int. Genet. 6 (2012) 387–392.
[17] J.-A. Bright, D. Taylor, J. Curran, J. Buckleton, Searching mixed DNA profiles directly against profile databases, Forensic Sci. Int. Genet. 9 (2014) 102–110.

Manuscript: Testing likelihood ratios produced from complex DNA profiles. D Taylor, J Buckleton, I Evett. (2015) Forensic Science International: Genetics 16, 165-171 – *Cited 8 times*

Statement of novelty: This work took existing theory from statistics and applied the theory to complex DNA profiling problems.

My contribution: Main author and sole simulation programmer. Equal contributor to theory.

Research Design / Data Collection / Writing and Editing = 60% / 100% / 33%

Additional comments:

Forensic Science International: Genetics 16 (2015) 165–171

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

Forensic population genetics – original research

Testing likelihood ratios produced from complex DNA profiles

Duncan Taylor [a,b,*], John Buckleton [c], Ian Evett [d]

[a] Forensic Science South Australia, 21 Divett Place, Adelaide, SA 5000, Australia
[b] School of Biological Sciences, Flinders University, GPO Box 2100, Adelaide, SA 5001, Australia
[c] ESR, Private Bag 92021, Auckland 1142, New Zealand
[d] Principal Forensic Services Ltd., London, UK

ARTICLE INFO

Article history:
Received 24 November 2014
Received in revised form 13 January 2015
Accepted 16 January 2015

Keywords:
DNA profile interpretation
Mixtures
Likelihood ratios
Performance tests

ABSTRACT

The performance of any model used to analyse DNA profile evidence should be tested using simulation, large scale validation studies based on ground-truth cases, or alignment with trends predicted by theory. We investigate a number of diagnostics to assess the performance of the model using H_d true tests. Of particular focus in this work is the proportion of comparisons to non-contributors that yield a likelihood ratio (LR) higher than or equal to the likelihood ratio of a known contributor (LR_{POI}), designated as p, and the average LR for H_d true tests. Theory predicts that p should always be less than or equal to $1/LR_{POI}$ and hence the observation of this in any particular case is of limited use. A better diagnostic is the average LR for H_d true which should be near to 1. We test the performance of a continuous interpretation model on nine DNA profiles of varying quality and complexity and verify the theoretical expectations.

1. Introduction

Methods for evaluating DNA profiles have benefitted from recent improvements in modelling and software [1–7] which have enabled the generation of quantitative evidential weights from profiles that would hitherto have been considered uninterpretable [1,5,7,8]. There is a considerable need to evaluate the performance of interpretation software, and to calibrate the evidential weights that different packages produce. Doing so can be a difficult task, particularly when the software is designed to analyse profiles that require calculations beyond the reasonable ability of an analyst. In these instances the performance of the system itself must be based on simulation, large scale validation studies based on ground-truth cases, or alignment with trends predicted by theory. It will often require all three.

Recently the idea of performance tests on a per case basis has been advanced [9]. The suggestion is that a great many profiles are simulated and tested in the place of the person of interest (POI). The fraction of those tests producing a likelihood ratio (LR) greater than the LR for the POI (LR_{POI}) may be reported and it has been suggested that this could be interpreted as a p-value. We would prefer to avoid the mixing of terms from the Bayesian and frequentist methods and will call this fraction p but not interpret it as a p-value, which would be more familiar if we were testing some hypothesis. This value of p is relative to a simulation and a model and care should be taken when moving this inference to a real population.

LRs produced for non-contributors that support their presence in a DNA sample have classically been referred to as 'adventitious matches' [10]. This term uses binary interpretation terminology and would perhaps better be referred to as 'misleading LRs' in a continuous framework as a 'match' loses its meaning when no single contributor's genotype can be resolved.

There is a limit in the type of DNA profiles that can be assessed by simulation because calibrating an LR of x requires simulation that has many more than x elements. For a complete DNA profile in a modern profiling system the value of x can be over 20 orders of magnitude, which is well beyond the practical limits of any standard computer.

Dørum et al. [11] outline a method to overcome this restriction. The fraction of genotypes equalling or exceeding the LR is calculated exactly and assuming between locus independence.

The assumption of no between locus effects [11], however convenient, ignores the fact that the larger peaks at each locus are more likely to come from the same contributor. This assumption ignores valid information and would not be sustainable if some aspect of the calculation invokes distinguishable contributor orders (for an explanation of the contributor concept see [12]), for example, using different drop-out probabilities for different

* Corresponding author at: Forensic Science South Australia, 21 Divett Place, Adelaide, SA 5000, Australia. Tel.: +61 8 8226 7700; fax: +61 8 8226 7777.
E-mail address: Duncan.Taylor@sa.gov.au (D. Taylor).

http://dx.doi.org/10.1016/j.fsigen.2015.01.008
1872-4973/

contributors. It should therefore be considered a convenient but inefficient assumption and should be avoided when testing any advanced system that correctly treats between locus effects.

Correctly avoiding such assumptions means that testing any large LR requires an unreasonably large number of simulations and is unlikely to be achievable on a per case basis.

The tests simulating the situation where the POI is not a donor are more appropriately called 'H_d true' tests rather than performance tests. Good [13] (quoting Turing) stated "*the expected factor for a wrong hypothesis in virtue of any experiment is 1.*" The proof is given in Good and reproduced here:

$$\sum_i P(E_i|\overline{H})\frac{P(E_i|H)}{P(E_i|\overline{H})} = \sum_i P(E_i|H) = P(E_1 \cup E_2 \cup \ldots \cup E_n|H) = 1$$

where E_i is the ith possible outcome, H and $\overline{H}$ are two propositions. Placing this into the DNA context we write the evidence as the crime stain genotype, G_S, which is fixed, and the potential donor's genotype, G_i, which is varied,

$$\sum_i P(G_s, G_i|\overline{H})\frac{P(G_s, G_i|H)}{P(G_s, G_i|\overline{H})} = 1$$

Paraphrasing this equation: the average LR for the H_d true tests should be one. It is relatively easy to "see" this equation in action. Consider a single source stain giving a full profile, G_s. The probability of G_s is 1 in a billion. The LR for a POI containing the same alleles as G_s is 1 billion for the propositions

H_p: POI is the donor
H_d: an unknown person is the donor

It follows that if people were chosen at random (or simulated) we would expect that 1 in every billion people would have the same reference profile as the POI and yield an LR of 1 billion when compared to the evidence. We would also expect that 999,999,999 out of 1 billion randomly chosen people would have a different reference to the POI and yield an LR of 0 when compared to the evidence. The average LR of these 1 billion comparisons is then 1.

Consider now a more complex evidence profile and n simulated reference profiles that are compared with it. From the theory above we would expect the average LR of these n simulations to tend to 1 if the simulation and the model used to interpret the simulation are the same.

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{LR}_i \approx 1$$

where LR_i is the LR generated from the i^{th} simulation. We emphasise that this experiment is testing the mixture interpretation portion of the software since the simulation is creating an idealised population that is in Hardy–Weinberg and linkage equilibrium.

We are interested in the proportion of n simulations that would yield an LR greater than or equal to $\mathrm{LR_{POI}}$ which we term p. If we order the simulations in increasing order of magnitude and let LR_{n-k} be the first LR that equals or exceeds $\mathrm{LR_{POI}}$ then:

$$\frac{1}{n}\sum_{i=1}^{n-k-1} \mathrm{LR}_i + \frac{1}{n}\sum_{i=n-k}^{n} \mathrm{LR}_i \approx 1$$

$$\text{defining} \quad \delta = (1/n)\textstyle\sum_{i=1}^{n-k-1}\mathrm{LR}_i \quad \text{gives} \quad (1/n)\textstyle\sum_{i=n-k}^{n}\mathrm{LR}_i \approx 1-\delta,$$

where δ is non-negative.

Because $\mathrm{LR}_i \geq \mathrm{LR_{POI}}$ for all i from n–k to n we can say that $(1/n)\sum_{i=n-k}^{n}\mathrm{LR}_i = (k/n)\mathrm{LR_{POI}} + \varepsilon$ and $(k/n)\mathrm{LR_{POI}} = 1-\delta-\varepsilon$, where ε is also non-negative.

We define $p = k/n$, which is the proportion of simulations producing an LR greater than or equal to LR_{POI}. This is akin to Dørum et al's p-value but we write p to emphasise that this is a proportion from a simulated sample in our case.

We write the expected value of p as $\tilde{p}$ and suggest that this expectation is

$$\tilde{p} \approx \frac{1-\delta-\varepsilon}{\mathrm{LR_{POI}}} \tag{1}$$

Which, as both δ and ε are non-negative values, provides the inequality:

$$\tilde{p} \leq \frac{1}{\mathrm{LR_{POI}}} \tag{2}$$

$\delta = 0$ occurs when all profiles that do not align with that of the POI yield an LR of 0 (i.e. when there are no ambiguity in potential contributor genotypes) and $\varepsilon = 0$ (i.e. when i contains a single element) occurs when the POI has a profile that yields the maximum possible LR. If both δ and ε are 0 then $\tilde{p} = 1/\mathrm{LR_{POI}}$, which represents the maximum value that $\tilde{p}$ can ever take.

The derivation of Eqs. (1) and (2) did not require any assumption about the population and as such would allow a statement of p, if desired, to be made in every case provided that the software was producing the LR in a reasonable manner.

If the software is not producing the LR in a reasonable manner then a p term from a simulation will not recover the situation although it may alert us to the fact that the software is underperforming. In any case the value will be relative to an idealised population.

Verbalising Eq. (2): *The probability of observing a likelihood ratio of LR_{POI} or larger from an unrelated non-donor is less than or equal to 1 in LR_{POI}*

This gives a frequentist sounding interpretation to the LR but is actually a statement that follows from the laws of probability. It avoids the awkward interpretation of results of H_d true trials as a p-value. This suggests that a more viable route to case specific reinterpretations of the LR is to assess if the software is preforming in a reasonable manner in large scale H_d true tests such as advocated by Evett et al. in several forensic fields [14].

We extend Gill and Haned [9] here. If the software is accepted as performing appropriately then Eqs. (1) and (2) should apply in each case. Our criteria would be that the average H_d true LR should be close to 1 and that Eq. (2) should hold. If the average LR is greater than 1 then the software is on average non-conservative relative to the simulation. If it is less than 1 it is on average conservative.

It has been suggested that: "*These tests are used to evaluate the LR and provide an important indication that the reported statistic has meaning on a per-case basis. Indeed, the argument can be taken further since there is no reason why the performance test itself could not be used instead of the LR statistic. But this debate is reserved to future work.*" [9] and we hope that the work carried out here is a start to that debate.

If the $\mathrm{LR_{POI}}$ is 1000 but $p = 1$ in a million then a suggested inference has been that the LR is very robust [11,15] and probably overly conservative. This would indeed lead to a value for p that is much smaller than $1/\mathrm{LR_{POI}}$.

However, given Eq. (1), we would suggest that such a value for p might instead suggest that δ or ε or both are large and the LR is not too small. It might suggest that $\delta + \varepsilon = 0.999$ in the above example. In fact, reflecting on Eq. (1) we suggest that it is not reasonable to interpret the value of p numbers as supporting the robustness of the LR.

We note here that hypotheses can be constructed so that Eq. (2) will not hold. Consider a two person profile made up from G_1 and G_2. If we simulated random profiles, G_R, the propositions:

H_{p1}: G_{R1} and G_{R2} are the donors
H_{d1}: Two unknown persons, unrelated to G_{R1} or G_{R2} are the donors

then the average LR from numerous simulated comparisons (where G_{R1} and G_{R2} are both simulated at every iteration) would be 1. If, however, the following propositions were addressed:

H_{p2}: G_1 and G_{R1} are the donors
H_{d2}: Two unknown persons, unrelated to G_1 or G_{R1} are the donors

then the average LR from numerous simulated comparisons would not equal 1. Again this can be 'seen' by considering a simple example. Let the probability that a person from the relevant population would be genotype G_1 be 1 in 1 million and let the probability of genotype G_2 be 1 in 1000. The DNA of G_1 and G_2 are combined to produce a mixed DNA profile with completely resolvable components. Simulations are carried out randomly generating DNA profiles and calculating an LR for H_{p2} and H_{d2}. On average, in every 1000 simulations we would expect 999 simulated profiles to differ from G_2 and so give an LR of 0 and 1 simulated profile to be the same as G_2 and give an LR of 1 billion. The average of these LRs is therefore 1 million.

The reason for the divergence from the expectations of Turing is that this analysis does not address propositions that represent H_d as the ground truth. Under H_{p2},G_1 and G_{R1} are the donors. This is clearly not reflective of the ground truth as G_{R1} is a simulated non-donor to the profile. Under H_{d2} the donors are two unknown individuals, unrelated to G_1 and G_{R1}. This also does not represent a ground truth as in this instance G_1 has specifically been identified as a non-contributor, which is not the case. For the expectations of Turing to hold, all known contributors being considered in the calculation must be present as donors in both H_p and H_d, and all simulated contributors must be present as donors in H_p and as non-donors in H_d.

For our two person example, acceptable sets of propositions for H_d are:

H_{p3}: G_1 and G_{R1} are the donors
H_{d3}: G_1 and an unknown person, unrelated to G_{R1} are the donors
H_{p4}: G_2 and G_{R1} are the donors
H_{d4}: G_2 and an unknown person, unrelated to G_{R1} are the donors
H_{p5}: G_{R1} and an unknown are the donors
H_{d5}: Two unknown persons, unrelated to G_{R1} are the donors
H_{p1}: G_{R1} and G_{R2} are the donors
H_{d1}: Two unknown persons, unrelated to G_{R1} or G_{R2} are the donors

We test the performance of the profile interpretation STRmix™ [16] against these criteria.

2. Method

Three constructed DNA samples were profiled using GlobalFiler™ (Thermo Fisher Scientific) as per manufacturer's instructions and six constructed DNA samples were profiled using Profiler Plus™ (Thermo Fisher Scientific) as per manufacturer's instructions, except at half the volume. DNA amounts and mixture proportions are given in Table 1 for the constructed DNA samples. Amplification fragments were resolved using the ABI PRISM® 3130xl Genetic Analyser and analysed in Genemapper® ID-X or Genemapper® ID 3.2.1.

DNA profiles were analysed using STRmix™ version 2.3 (http://strmix.esr.cri.nz). Following Gill and Haned [9] random DNA profiles were generated in proportions according to expectation from allele frequencies. This effectively simulates a population in Hardy–Weinberg and linkage equilibrium. We emphasise that this does not replicate a real population but rather an idealised one. Hence this test does not validate the full software performance either for STRmix™ or the previously reported work on LRmix [9]. It only calibrates that portion of the software separate from the population genetic model.

These randomly generated DNA profiles were compared to the mixed DNA profiles to generate LRs. All LRs were calculated using Caucasian databases (either an in-house database for GlobalFiler profiles or [17] for Profiler Plus calculations) and using the product rule. By using the product rule to calculate LRs we align the method used to simulate the profiles with the method used to interpret them. Simulation using the product rule and interpreting using the Balding and Nichols equations [18] creates an inconsistency and hence it is very hard to predict the correct result. However for a discrete system where $\delta = \varepsilon = 0$ this discrepancy would force $p \leq 1/\mathrm{LR}_{\mathrm{POI}}$ and the average LR for H_d true to be less than 1 inappropriately.

Table 1
Experimental setup, an asterisk (*) denotes a sample where loci containing STR information have been ignored. Instances where the value in the 'STR loci' column is less than the number of autosomal STR loci in the kit, in the absence of an asterisk, indicates a partial evidence profile was obtained.

Experiment	Profiling system	Contributors	STR loci	Known contributor DNA amounts (pg)	Propositions: C = known contributor U = unknown RANDOM = simulated profile
1	GlobalFiler	4	21	13:13:13:13	H_p: C2 + C3 + C4 + RANDOM H_d: C2 + C3 + C4 + U
2	GlobalFiler	4	21	20:15:10:5	H_p: C1 + C2 + C3 + RANDOM H_d: C1 + C2 + C3 + U
3	GlobalFiler	4	12	4:3:2:1	H_p: RANDOM + 3U H_d: 4U
4	Profiler Plus	1	6	20	H_p: RANDOM H_d: U
5	Profiler Plus	1	3	10	H_p: RANDOM H_d: U
6	Profiler Plus	2	6*	38:38	H_p: RANDOM + U H_d: 2U
7	Profiler Plus	3	9	100:100:100	H_p: RANDOM + 2U H_d: 3U
8	Profiler Plus	1	4*	500	H_p: RANDOM H_d: U
9	Profiler Plus	1	1*	500	H_p: RANDOM H_d: U

Table 2
Results of comparisons of simulated random references to profiles outlined in Table 1. Vertical axes for distributions show counts.

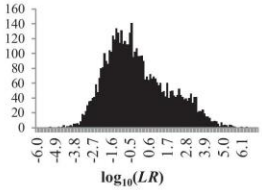
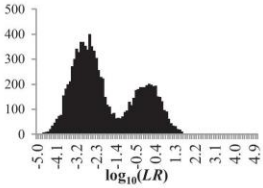
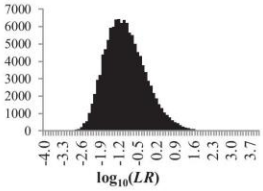
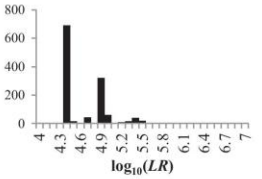
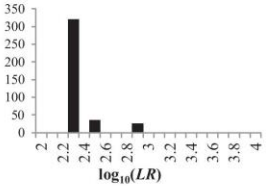
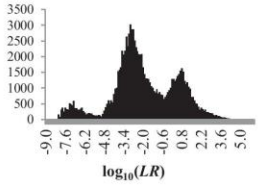
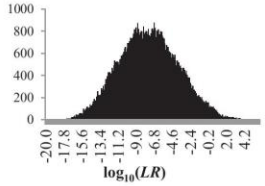
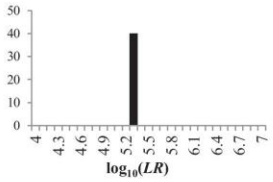
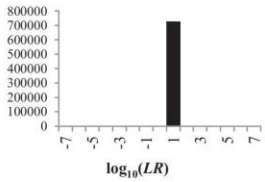
Experiment	Simulations	H_d true $\log_{10}$(LR)	Number of H_d true with LR = 0	Average H_d true LR	Counter number	H_p true LR (s)	p, '1 in'
1	12,000,000		11,994,959	1.0046	C_4	3,74,104	3,000,000
2	10,000		0	0.977	C_4	9	44
3	1,20,000		0	0.927	C_1 C_2 C_3 C_4	4 7 5 6	29 56 34 49
4	80,000,000		922,585	1.001	C_1	3,12,325	6,666,666
5	100,000		99,618	1.022	C_1	215	262
6	10,000,000		9,898,155	1.017	C_1 C_2	278 12,557	4,163 78,125

Table 2 (*Continued*)

Experiment	Simulations	H_d true $\log_{10}$(LR)	Number of H_d true with LR = 0	Average H_d true LR	Counter number	H_p true LR (s)	p, '1 in'
							
7	1,000,000		922,585	0.906	C_1 C_2 C_3	2,34,738 2,530 43	>1,000,000 17,241 2,262
8	10,000,000		9,999,960	0.872	C_1	218,070	250,000
9	10,000,000		9,274,620	1.003	C_1	14	14

By aligning the two we expect the average LR for H_d true to be close to 1 and recognise explicitly that we are testing the mixture interpretation and not the population genetic model. Table 1 shows the experimental setup and the propositions tested.

In order to make the problem tractable we have reduced the number of loci and hence the LR in some cases (as indicated in Table 1). Specific details regarding the STRmix™ settings used to analyse the DNA profiles are available on request from the authors.

3. Results

LRs generated in the comparisons of randomly simulated profiles to the evidence profiles were used to generate the distributions seen in Table 2. Also shown in Table 2 is the average H_d true LR value from all simulations, the LRs for the known contributors and p expressed as '1 in' for each of the known contributor's LRs. The value of p has been given as '1 in' for ease of direct comparison to the LR values of known contributors, in the check for adherence to Eq. (2). In short, for Eq. (2) to hold then $1/p$ must be larger than or equal to the LR of the known contributor it is being compared against.

We recognise that in some laboratories the number of unknown contributors under H_p is increased if the person hypothesised as a contributor is excluded. This was not done in our study.

Experiment 1. A complex four person profile constructed with equal mixture proportions and assuming three of the four known contributors. The H_p true LR was 374,104 and from this information the expectation was that the number of H_d true LR = 0 would be high. This is indeed what was observed. The value of p was 1 in 3 million, which is almost an order of magnitude below LR_{POI} and so Eq. (2) holds. The average H_d true LR is 1.0046, which is very close to 1 as expected from theory.

Experiment 2. The experiment was similar in setup to experiment 1, however the contributor proportions in the mixture were unequal. Again three of the four known contributors were assumed, with the unassumed contributor being the minor contributor to the profile. An H_p true LR of 9 was

obtained which indicated that the amount of information in the unassumed contributor position was low. This is backed up by the fact that none of the simulated profiles gave an H_d true LR = 0. A profile setup in this manner is not subject to high sampling variation in simulation numbers as it is not relying on observing a very rare event. This meant the number of H_d true simulations could be kept down to 10,000 and still show an average H_d true LR of close to 1 (0.977). Again Eq. (2) holds with $1/p \sim 44$.

Experiment 3. This experiment investigates a low level four person mixture where none of contributors are assumed. All H_p true LRs were low and again there were no instances of H_d true LR = 0. The average H_d true LR was 0.927, and Eq. (2) held for all H_p true LRs.

Experiment 4. This experiment investigated a single source profile with six loci of which only one was able to be unambiguously assigned as a heterozygote. All other loci contained a single peak and had both heterozygote (drop-out) and homozygote (non drop-out) genotypes assigned to them with non-zero weight. The known contributor aligned with the heterozygote (drop-out) genotype in three of the five ambiguous loci. The limited number of different H_d true LRs reflects the restricted number of genotypes that randomly generated profiles can have that will not lead to an H_d true LR = 0. Again Eq. (2) held for this example and the average H_d true LR was close to 1.

Experiment 5. This experiment contained three loci, all of which contained a single, low level peak, that could be explained by contribution of DNA from either heterozygote or homozygote sources. This experiment is similar to experiment 4, just with fewer loci, and lower peak heights.

Experiment 6. This experiment examined a two person mixed DNA profile. Three loci were ignored so that the number of simulations could be kept down to 10 million. The individuals who contributed were intended to be added in even amounts, however one contributor (C_1) was slightly less intense that the other (C_2) and this is reflected by the H_p true LRs. Eq. (2) held true and the average H_d true LR was close to 1. Because of the level of discrimination in the profile there were a large number of H_d true LR = 0.

Experiment 7. This profile was generated from three individuals, who contained a lot of masking. Only two of the nine STR loci exhibited more than four allelic peaks. The result was a range of H_p true LRs. Eq. (2) held true, with no observations of an H_d true LR appearing above the H_p true LR for C_1. Again the average H_d true LR was close to 1.

Experiment 8. This profile was a complete single source profile. For the analyses only four loci were included so that simulations could be kept to 10 million. In this instance we would expect $1/p = LR_{POI}$ and indeed the two values were close. The slight divergence of these two values is likely due to sampling variation in the simulation process and this is reinforced by the average H_d true LR value being slightly lower than the expected value of 1. To further investigate that this was the case experiment 9 was conducted. Note that being a strong single source profile there is only one genotype that will yield a non-zero LR. This genotype will always give the same LR (due to the simplifications of the model we are using) and so only a single bar is seen on the distribution of H_d true LRs for this sample. This is also true of experiment 9.

Experiment 9. This profile was again a complete single source profile, however only a single unambiguous locus was used for simulation. The H_p true LR was 14 and as expected by theory in this situation $1/p = 14$ also. The average H_d true LR was very close to one.

Table 2 shows that the H_d true log(LR) distributions can be multi-modal. We believe that this may be due to 'groups' of genotype sets that share certain properties e.g. levels of homozygosity, number of drop-outs, number of drop-ins or certain peak imbalances. The interaction that leads to these groupings is likely to be complex and we have not attempted to investigate their source as it is a side issue to the focus of this work.

4. Discussion

In all cases tested Eq. (2) held and the average H_d true value is close to one. The fact that $1/p$ is larger than LR_{POI} does not provide an indication that the LRs produced for POIs are robust, but rather is an expected outcome from probability theory for complex profiles. The closeness of the average of H_d true LRs to one is a better indication of the robustness of the LR and even then we reiterate that this is more an indication of the performance of the models used to generate the LR than of the LR itself. Recall that this test, as with the performance tests of Gill and Haned [9], are relative to a population generated under the product rule expectations. However no assumption of Hardy–Weinberg and linkage equilibrium is made in deriving Eq. (2) and hence this equation is expected to hold for real populations if the LR is assigned appropriately.

If the work described here is correct then the idea that the result of a performance test could replace the LR cannot be supported. The probability, p, differs from $1/LR_{POI}$ because of the expected properties of probability and does not add to the LR. We do not suggest that the LR be replaced by this statement but rather suggest that the LR is the most informative statistic possible. Further we suggest that the probability p might be misleading if, for example, the LR is 1000 and $p = 1$ in a million. If the LR is a reasonable assignment of the weight of evidence then the p term may lead to an overestimation of the evidence in some triers of fact.

The results of this work also go towards addressing a misunderstanding that has arisen recently during defence expert testimony. This questions the reliability of an LR, when it is derived from low quantities of DNA. We provide two examples of defence expert testimony as example [R v FULLER, District Court of South Australia, August 2014]. We have added information enclosed in square brackets, to which we assume the witness is referring, in order to make the statements more comprehensible.

"*We have an issue where we have got a little bit of DNA of someone you might be really interested in, then the software, based on validation work the laboratory has done, is very poor at measuring that difference* [between contributors and non-contributors] *because there is such a high level* [of peak height variability]"

And later in the same trial:

"*I say, based on what their published validation stuff says, the spread* [of heterozygous balance] *when you get down low is so great that it is impossible for the software to accurately assess it, or predict it* [supported genotypes] *I should say.*"

The findings of this study further demonstrate a point raised in [16], which is that regardless of the strength or complexity of the DNA data, as long as the models used to analyse the data are reliable, then the LR produced will also be reliable. In that work, a number of deliberately low level DNA profiles were chosen (both single source and mixed) to demonstrate this concept. A low LR for the comparison of a reference to such samples, which is the type of result commonly

referred to as 'unreliable', is simply demonstrating the expected (and desired) behaviour of LRs.

We recognise here that the reality of casework is that there is generally a complexity 'threshold' where DNA profiles will not be analysed. This threshold should not be taken as evidence of the non-reliability of a model's performance, but rather a practical or business decision made by the analyst.

It is a natural concern that is sometimes expressed by forensic practitioners that, where profiles contain relatively small peaks, with consequent appreciable levels of uncertainty in their measurement and designation, then the LR from a continuous calculation becomes increasingly unreliable. However, provided the modelling assumptions have been adequately informed by experimental data then the increased levels of uncertainty in low level profiles lead to LR distributions that increasingly trend towards one, adequately and coherently reflecting the decrease in information content of such profiles. Experiments of the kind reported in this paper are able to demonstrate this.

Rather than a reduction in reliability, the LR trending towards one as the amount of information decreases, is a drop in the informativeness of the result, i.e. the ability of the result to inform us of some support for one proposition over another. In this instance the probability that a randomly chosen non-donor could yield the same (or higher) LR is increased, however the size of the LR is still the best reflection of this fact. Somewhat counterintuitively it is the simplest of profiles (unambiguous and single sourced) where the probability of a randomly chosen non-donor giving the same (or higher) LR will most nearly approach $1/LR_{POI}$, and not the complex or low level profiles.

The adherence of results in this work to the concept "*the expected factor for a wrong hypothesis in virtue of any experiment is 1*", even at low levels, is yet another demonstration of the reliability of the LR statistic for analysis of DNA profile evidence, when generated with reliable models.

Acknowledgements

This work was supported in part by grant 2011-DN-BX-K541 from the US National Institute of Justice. Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the U.S. Department of Justice. We thank Catherine McGoven, Stuart Cooper and two anonymous referees for comments that have greatly improved this paper.

References

[1] D. Taylor, J.-A. Bright, J. Buckleton, The interpretation of single source and mixed DNA profiles, Forensic Sci. Int. Genet. 7 (2013) 516–528.
[2] H.H. Forensim, An open-source initiative for the evaluation of statistical methods in forensic genetics, Forensic Sci. Int. Genet. 5 (2011) 265–268.
[3] K. Lohmueller, N. Rudin, Calculating the weight of evidence in low-template forensic DNA casework, J. Forensic Sci. 58 (2013) 234–259.
[4] C.D. Steele, D.J. Balding, Statistical Evaluation of Forensic DNA Profile Evidence, Ann. Rev. Stat. Appl. 1 (2014) 361–384.
[5] R. Puch-Solis, L. Rodgers, A. Mazumder, S. Pope, I. Evett, J. Curran, et al., Evaluating forensic DNA profiles using peak heights, allowing for multiple donors, allelic dropout and stutters, Forensic Sci. Int. Genet. 7 (2013) 555–563.
[6] D.J. Balding, J. Buckleton, Interpreting low template DNA profiles, Forensic Sci. Int. Genet. 4 (2009) 1–10.
[7] M.W. Perlin, M.M. Legler, C.E. Spencer, J.L. Smith, W.P. Allan, J.L. Belrose, et al., Validating TrueAllele® DNA mixture interpretation, J. Forensic Sci. 56 (2011) 1430–1447.
[8] R.G. Cowell, S.L. Lauritzen, J. Mortera, Probabilistic modelling for DNA mixture analysis, Forensic Sci. Int. Genet. 1 (2008) 640–642 Supplement series.
[9] P. Gill, H. Haned, A new methodological framework to interpret complex DNA profiles using likelihood ratios, Forensic Sci. Int. Genet. 7 (2013) 251–263.
[10] P.M. Schneider, DNA databases for offender identification in Europe - the need for technical, legal and political harmonisation, The Second European Symposium on Human Identification, Innsbruck, Austria, 1998.
[11] G. Dørum, Ø. Bleka, P. Gill, H. Haned, L. Snipen, S. Sæbø, et al., Exact computation of the distribution of likelihood ratios with forensic applications, Forensic Sci. Int. Genet. 9 (2014) 93–101.
[12] D.A. Taylor, J.-A. Bright, J.S. Buckleton, The 'factor of two' issue in mixed DNA profiles, J. Theor. Biol. 363 (2014) 300–306.
[13] I.J. Good, Probability and the Weighing of Evidence, Charles Griffin & Company Limited, London, 1950.
[14] I.W. Evett, J.A. Lambert, J.S. Buckleton, B.S. Weir, Statistical analysis of a large file of STR profiles of British Caucasians to support forensic casework, Int. J. Legal Med. 109 (1996) 173–177.
[15] L. Prieto, H. Haned, A. Mosquera, M. Crespillo, M. Alemañ, M. Aler, et al., Euroforgen-NoE collaborative exercise on LRmix to demonstrate standardization of the interpretation of complex DNA profiles, Forensic Sci. Int. Genet. 9 (2014) 47–54.
[16] D. Taylor, Using continuous DNA interpretation methods to revisit likelihood ratio behaviour, Forensic Sci. Int. Genet. 11 (2014) 144–153.
[17] S.J. Walsh, J.S. Buckleton, Autosomal microsatellite allele frequencies for a nationwide dataset from the Australian Caucasian sub-population, Forensic Sci. Int. Genet. 168 (2007) e47–e50.
[18] D.J. Balding, R.A. Nichols, DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands, Forensic Sci. Int. Genet. 64 (1994) 125–140.

5a – clarification

The specifications of the computer used to carry out the simulations in the previous paper are:

- Intel(R) Core(TM) i7-3940XM CPU@3.00GHz
- 32 GB RAM
- 64-bit Windows 7 Ultimate

Manuscript: Importance sampling allows Hd true tests of highly discriminating DNA profiles. D Taylor, J Curran, J Buckleton. (2017) Forensic Science International: Genetics – *accepted, in press*

Statement of novelty: This work extends the work from the previous paper. Again, it took existing theory from statistics (in this case importance sampling) and demonstrated the application of the theory to complex DNA profiling problems.

My contribution: Main author and sole simulation programmer. Equal contributor to theory.

Research Design / Data Collection / Writing and Editing = 33% / 100% / 70%

Additional comments: During publication equations 3, 4 and 5 were incorrectly formatted by the journal. They should be:

$$ {}^{n}w_T^l = \sum_{j=1}^{J^l} \begin{cases} w_j^l & {}^{n}G_j^l = G_T^l \\ 0 & {}^{n}G_j^l \neq G_T^l \end{cases} \qquad (3) $$

Noting that now weights for each contributor at a locus will sum to one, $\sum_{j=1}^{{}^{n}J^l} {}^{n}w_j^l = 1$.

$$ b = \frac{f(P)}{(N-M)^{-1}\prod_l {}^{n}w_j^l f(Q^l)} \qquad j \in \mathbb{Z} : j \in \left\{1 \ldots {}^{n}J^l\right\} \qquad (4) $$

$$ b = \frac{f(P)}{\sum_{x=1}^{N-M} \frac{1}{N-M} \prod_l {}^{x}w_T^l f(Q^l)} \qquad (5) $$

Forensic Science International: Genetics 27 (2017) 74–81

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

Research paper

Importance sampling allows H_d true tests of highly discriminating DNA profiles

Duncan Taylor[a,b,*], James M. Curran[c], John Buckleton[d,e]

[a] Forensic Science South Australia, 21 Divett Place, Adelaide, SA 5000, Australia
[b] School of Biological Sciences, Flinders University, GPO Box 2100 Adelaide SA, Australia 5001 ESR, Private Bag 92021, Auckland 1142, New Zealand
[c] Department of Statistics, University of Auckland, Private Bag 92019, Auckland 1142, New Zealand
[d] National Institute of Standards and Technology, 100 Bureau Drive, MS 8980 and 8314, Gaithersburg, MD 20899, United States
[e] ESR, Private Bag 92021, Auckland 1142, New Zealand

ARTICLE INFO

Article history:
Received 21 July 2016
Received in revised form 1 December 2016
Accepted 8 December 2016
Available online 9 December 2016

Keywords:
DNA profile
H_d true
Likelihood ratios
Performance tests
Importance sampling

ABSTRACT

H_d true testing is a way of assessing the performance of a model, or DNA profile interpretation system. These tests involve simulating DNA profiles of non-donors to a DNA mixture and calculating a likelihood ratio (*LR*) with one proposition postulating their contribution and the alternative postulating their non-contribution. Following Turing it is possible to predict that "*The average LR for the H_d true tests should be one*" [1]. This suggests a way of validating softwares. During discussions on the ISFG software validation guidelines [2] it was argued by some that this prediction had not been sufficiently examined experimentally to serve as a criterion for validation. More recently a high profile report [3] has emphasised large scale empirical examination.

A limitation with H_d true tests, when non-donor profiles are generated at random (or in accordance with expectation from allele frequencies), is that the number of tests required depends on the discrimination power of the evidence profile. If the H_d true tests are to fully explore the genotype space that yields non-zero *LR*s then the number of simulations required could be in the 10 s of orders of magnitude (well outside practical computing limits). We describe here the use of importance sampling, which allows the simulation of rare events to occur more commonly than they would at random, and then adjusting for this bias at the end of the simulation in order to recover all diagnostic values of interest. Importance sampling, whilst having been employed by others for H_d true tests, is largely unknown in forensic genetics. We take time in this paper to explain how importance sampling works, the advantages of using it and its application to H_d true tests. We conclude by showing that employing an importance sampling scheme brings H_d true testing ability to all profiles, regardless of discrimination power.

1. Introduction

A recent publication [1] examined a method of simulation-based performance testing of a model [4,5] used to evaluate DNA profiling data. These tests involved simulating DNA profiles of non-donors to a DNA mixture and calculating a likelihood ratio (*LR*) with one proposition postulating their contribution and the alternative postulating their non-contribution. Tests simulating the situation where a person of interest (POI) is not a DNA donor are more appropriately called 'H_d true' tests rather than performance tests. Good [6] (quoting Turing) stated "*the expected factor for a wrong hypothesis in virtue of any experiment is 1.*" Focussing this to the problem at hand translates to "*The average LR for the H_d true tests should be one*". In [1] the truth of this lemma was demonstrated by the use of H_d true tests on nine DNA profiles of varying complexity and information content This suggests a way of validating softwares by noting the average *LR* in a large number of Hd true tests.[1] During discussions on the ISFG software validation

* Corresponding author at: Forensic Science South Australia, 21 Divett Place, Adelaide, SA 5000, Australia.
E-mail address: Duncan.Taylor@sa.gov.au (D. Taylor).

http://dx.doi.org/10.1016/j.fsigen.2016.12.004
1872-4973/

[1] Note that adherence to this lemma is not the only test that a system would need to pass in order to be considered valid. The adherence of a system to the lemma follows from the laws of probability, hence while it will demonstrate that a probability distribution has been formed on the genotypes it does not mean that the probability distribution is sensible. Secondly, we do not know how systems will behave that treat nuisance parameters in the model differently under Hp and Hd, but it is quite probable that they will not adhere to the lemma. The implications of this behaviour are beyond the scope of this article.

guidelines [2] it was argued by some that this prediction had not been sufficiently examined experimentally to serve as a criterion for validation. More recently a high profile report [3] has emphasised large scale empirical examination.

The discrimination power of the DNA profiles that can be tested by standard sampling is limited. An *LR* of *x* requires simulation that has many more than *x* elements. For a complete DNA profile in a modern profiling system the value of *x* can be over 20 orders of magnitude, which is well beyond the practical limits of any standard computer using a naïve simulator.

In the simulations carried out in [1] on profiles with highly discriminating information the vast majority of *LR*s produced had a value of zero. A situation can be imagined where a single source DNA profile that had a profile frequency of 1 in 1 billion was undergoing H_d true tests using propositions:

H_p The randomly simulated non-donor is the source of the DNA
H_d An unknown individual is the source of the DNA

Within each block of one billion tests we would expect an *LR* of one billion to be obtained once, and the rest of the simulations would yield an *LR* of zero. Most observers would agree that this seems like a large effort to obtain mostly zeros. A more efficient system would simulate profiles that we knew were not going to yield an *LR* of zero, and as long as we knew what proportion would give zero (had we carried out the naïve simulation) then we would end up with the same total information as using a naïve simulator. The advantage, however, is a very much reduced requirement for simulation. In the single source example described above, we would only need to run one test that simulated the one profile that gave an inclusionary *LR* and as long as we know that the frequency of obtaining a non-zero *LR* was1×10^{-9}, then we would have the same information as before but at 1 billionth the computing cost. This is the idea behind a technique known as 'importance sampling'.

This idea is not new to forensic biology. In a recent publication describing the workings of continuous DNA interpretation software [7], importance sampling was used to consider genotypes that are included in an assessment of the probability of the evidence. In [8] the authors demonstrate the workings of importance sampling as applied to choosing genotypes to calculate the proportion of *LR*s derived from mixtures above a chosen value. Prior to this importance sampling was very nicely demonstrated in [9] with application to calculating exceedance probabilities. Despite these publications, the idea of importance sampling can be difficult to understand for those who do not have a statistical background. We attempt, in this work, to explain what importance sampling is, with simple examples, and how it is beneficial when using a sampling system to assign a probability for the occurrence of rare events.

We demonstrate the application of importance sampling to H_d true tests so that all profiles (of any discrimination capacity) are within the realms of being practically demonstrated to adhere to Turing's lemma [6]. This is an important ability to possess for model validation, particularly with regards to highly sophisticated DNA evidence interpretation systems.

2. Theory

Importance sampling biases the simulation process so that some elements are chosen more often than at random, and then readjusts for the bias after the simulation. The topic of importance sampling often arises in situations where we want to estimate the probability of a rare event. Importance sampling solves this problem by sampling from an importance density and reweighting the sampled observations accordingly. In general, if X is a random variable with probability density function$p(x)$, and $f(X)$ is some function of X, then the expected value of $f(X)$ is

$$\mathrm{E}[f(X)] = \int_{-\infty}^{+\infty} f(x)p(x)dx$$

If $h(x)$ is also a probability density function which is greater than or equal to zero for the same range of values as $p(x)$ (that is it lies within the support of $p(x)$), then this integral can be rewritten as

$$\mathrm{E}[f(X)] = \int_{-\infty}^{+\infty} f(x)\frac{p(x)}{h(x)}h(x)dx$$

This statement is not very interesting in itself. After all it is equivalent to multiplication by one. However, it is the "trick" which underlies important sampling. If we take a large sample of size S from the *importance density*$h(x)$, then this integral can be approximated by

$$\mathrm{E}[f(X)] \approx \frac{1}{S}\sum_{i=1}^{S} w_i f(x_i)$$

where $w_i = p(x_i)/h(x_i)$ are the *importance weights*. The idea behind importance sampling is that the importance density $h(x)$, can be easier to sample from than the original density, $p(x)$, and yields a low-variance estimate of the desired expectation. The choice of $h(x)$ is somewhat arbitrary, but does dictate the efficiency of the sampling scheme. The process of choosing a good importance distribution is known as *tuning* and can often be very difficult. One might think of this process over-sampling the events of interest, and then *down-weighting* or *biasing* the sample values with weights that reflect the relative probabilities of the events in the importance and original densities. We provide a simple example of importance sampling in Appendix A.

2.1. Application of importance sampling to H_d true tests

In the problem at hand we might regard X as the *LR*, and $f(X) = X$. That is f is the identity. For each of the 'y' H_d true tests carried out we calculate a weight, which we call a bias and denote b_y. Here, b_y reflects the size of the bias that leads to the choice in test y. In words, the bias term is the ratio of the probability of the choice using an unbiased method to the probability of that choice had the biasing method been employed. An approximation of the average *LR* (over the Y tests) that would have been obtained had a naïve simulator been used is then:

$$\overline{LR} = \frac{1}{Y}\sum_{y} LR_y b_y \tag{1}$$

and the number of simulations (I) that this would have required had a naïve simulator been used can be approximated by (see Appendix B for derivation):

$$I = \frac{\sum_{y} LR_y}{\overline{LR}} \tag{2}$$

In our single source example from earlier, imagine that we had run one H_d test. The probability of choosing the one genotype that would give a non-zero *LR* using the biased method is one, and this would yield an *LR* of one billion. The probability of choosing this genotype given the unbiased method is 1 in one billion and so $LR_1 = 1 \times 10^9$, $b_1 = 1 \times 10^{-9}$ and $\overline{LR} = \frac{1}{1}\left(1 \times 10^9\right)\left(1 \times 10^{-9}\right) = 1$. The approximate number of iterations that this corresponds to using a naïve simulator is $I = \frac{1 \times 10^9}{1} = 1 \times 10^9$. This is exactly aligned with our initial expectations, outlined in the introduction. In many instances $\overline{LR} \approx 1$, simplifying Eq. (2) to $I \approx \sum_{y} LR_y$.

All that is left then is to set up the method of importance sampling so that the bias terms can be calculated. We consider that a system used to deconvolute a DNA profile will result in a list of genotypes, each with an associated value (or weight as it is commonly called). This value is the probability of the observed data given a genotype set, and can either be generated from a combination of discrete probabilities of drop-in or dropout in a semi-continuous framework, or generated by integration across a number of nuisance parameters that describe DNA profile features. Table 1 shows the structure of an output from the deconvolution.

In Table 1 genotypes are denoted by a capital G, with left superscript denoting the contributor position within the genotype set, right superscript denoting the locus and right subscript denoting the genotype set. A genotype set (S_j^l) is then the vector of genotypes in set j, $S_j^l = \left\{{}^1G_j^l, \ldots, {}^NG_j^l\right\}$. Note that for each locus $\sum_{j=1}^{J^l} w_j^l = 1$. Within the genotypes for any specific contributor there can exist redundancies, i.e. imagine a two person profile with a completely resolved major and an unresolved minor. The result would be an output similar to Table 1 but the genotypes of contributor 1 (who we designate as the major) would be fixed, i.e. ${}^1G_1^l = {}^1G_2^l = {}^1G_j^l$, and only the genotypes of contributor 2 (the minor) would change. In order to calculate the choice bias of a particular genotype we need to know the probability of choosing specific genotypes and so the genotype set list in Table 1, needs to be broken down by contributor as well as locus. For contributor n, at locus l, we define ${}^nJ^l$ as the number of unique genotypes they can possess. The weights, now broken down per locus and per contributor, for a test genotype G_T^l, can be determined by:

$$ {}^nw_T^l = \sum_{j=1}^{J^l} \left\{ w_j^l \quad {}^nG_j^l = G_T^l 0 {}^nG_j^l \neq G_T^l \right. \tag{3}$$

Noting that now weights for each contributor at a locus will sum to one, $\sum_{j=1}^{n} J^l {}^n w_j^l = 1$.

In our biasing of genotype choice, in order to remove all instances of obtaining an *LR* of 0, we choose a contributor genotype held within the deconvoluted list, with probability according to its weight (i.e. a genotype with weight ${}^nw_T^l = 0.9$ will be chosen with probability 0.9). The probability of this choice has the following elements:

a A probability of $1/N$ of choosing any of the N contributor positions. Note that if M individuals have been assumed as contributors to the profile then the choice would be amongst the remaining, non-assumed, contributor positions and the probability would be $(N-M)^{-1}$

b A probability of $\prod_l {}^nw_j^l$ of choosing the genotype for that contributor, multiplied across all locus choices

c A probability $\prod_l f(Q^l)$ that is invoked to account for the genotype chosen that possesses a dropout allele, denoted Q (and defined as any allele other than those already able to be possessed by that contributor at the locus in question from the deconvolution list), multiplied across loci

Within the list of genotypes that a contributor can possess there are three possibilities at a locus:

a The genotype specifies both alleles explicitly, in which case $f(Q^l) = 1$, i.e. the bias in picking the genotype by the importance sampler does not need to consider an allele frequency

b The genotype possesses one Q allele, which must be replaced with an allele from the population from the available alleles (i.e. any allele other than those already able to be possessed by that contributor at the locus in question from the deconvolution list). If allele A is chosen then $f(Q^l) = \frac{p_A^l}{p_Q^l}$, where p_A^l is the probability of choosing allele A at locus l, in other words, the bias in picking the genotype by the importance sampler must take into account the frequency of the allele chosen

c The genotype possesses two Q alleles, in which case they must both be replaced from the available alleles and $f(Q^l) = (p_A^l)^2 / \left(p_Q^l\right)^2$ if the same allele, A, is chosen twice or $f(Q) = 2p_A^l p_B^l / \left(p_Q^l\right)^2$ if two different alleles, A and B, are chosen. In this scenario, the bias in picking the genotype by the importance sampler must take into account the frequency of the genotype.

The unbiased probability of choosing the genotype is simply the profile frequency (in our example calculated using assumption of Hardy-Weinberg equilibrium), which we represent by $f(P)$, and so the bias term for a particular choice of non-donor profile is:

$$ b = \frac{f(P)}{(N-M)^{-1} \prod_l^n w_j^l f(Q^l)} \quad \mathrm{j} \in \mathbb{Z} : \mathrm{j} \in \left\{1 \ldots {}^nJ^l\right\} \tag{4}$$

Again, as a demonstration, consider the single source profile in our running example, $N = 1$, $M = 0$. There are no dropouts and so $\prod_l f(Q^l) = 1$, and as there is only one genotype to choose at each locus, ${}^1w_1^1 = {}^1w_1^2 = {}^1w_1^l = 1$ and so $\prod_l {}^nw_j^l = 1$. The profile frequency is $f(P) = 1 \times 10^{-9}$ and so the bias term is:

$$ b = \frac{1 \times 10^{-9}}{(1-0)^{-1} 1 \times 1} = 1 \times 10^{-9} $$

as previously assigned when using intuition.

We need to extend the theory to calculate b in one additional way, which cannot be demonstrated with the running single source example. The method, as described above, makes the assumption that the two contributors in a deconvolution cannot

Table 1
Structural representation of deconvolution output of a profile at locus l, showing the J^l genotype sets.

Genotype set (j)	Genotype (G) of contributor 1	...	Genotype (G) of contributor N	Weight (w)
1	${}^1G_1^l$	...	${}^NG_1^l$	w_1^l
2	${}^1G_2^l$	...	${}^NG_2^l$	w_2^l
⋮	⋮	⋱	⋮	⋮
J^l	${}^1G_{J^l}^l$	...	${}^NG_{J^l}^l$	$w_{J^l}^l$

possess the same whole profile genotype i.e. once a contributor position has been chosen, it is assumed that no choice of full profile genotype is also able to be made in any other contributor. This is often not the case. Image a two person profile, where both contributors have donated equal amounts of DNA. The deconvolution of the mixture will give rise to a list of genotype sets that possess the same complement of genotypes for each contributor, and with equal per-contributor weights. In other words, if one of the two contributors was chosen, and a genotype chosen at each locus from those available to that contributor, then using Eq. (4) to calculate the bias would be a factor of two too low. The reason for this is that the chance of that genotype being chosen using our biased sampling method needs to take into account that the genotype could have been chosen if either contributor 1 or contributor 2 was the initial choice, and in Eq. (4), only the latter (or only the former) is considered. This contributor ordering is the same phenomenon that leads to the factor of $N!$ required to elevate sub-sub source level propositions to sub-source level propositions in DNA profile evidence evaluation [10].

We define the choice of genotype at locus l in contributor n as the target genotype, G_T^l, which in contributor x has per-contributor weight ${}^x w_T^l$. We seek the probability of choosing that genotype in each of the $N-M-1$ contributors that were not initially chosen. To simplify this mathematically, consider the probability of choosing a target genotype in any of the $N-M$ contributors. Bias can be calculated by:

$$b = \frac{f(P)}{\sum_{x=1}^{N-M} \frac{1}{N-M} \prod_l {}^x} w_T^l f(Q^l) \tag{5}$$

and the choice of genotypes is such that the reference profile will not yield $LR = 0$ when compared to the evidence profile. In practice this is achieved by choosing a random contributor and then a genotype for that contributor at each locus (as described initially), however the method of genotype choice is not required for Eq. (5) when the requirement of $LR > 0$ is explicitly stated. Note that in the example of an equal two-person profile with no assumed contributors, Eq. (5) simplifies to:

$$b = \frac{f(P)}{\prod_l {}^x} w_T^l f(Q^l)$$

where x can be either contributor 1 or 2 as they are both equivalent.

3. Method

We carry out H_d true tests on profiles outlined in Table 2, using the mathematics outlined in the theory section, and Eq. (5) to calculate b. All profiles were generated either using Profiler Plus™ or GlobalFiler™ (Thermo Fisher Scientific) as per manufacturer's instructions (except that half volume PCR reactions were used for Profiler Plus™). DNA amounts and mixture proportions are given in Table 2 for the DNA samples. Amplification fragments were resolved using the ABI PRISM® 3130xl Genetic Analyser and analysed in Genemapper® ID-X. DNA profiles were analysed using STRmix™ [4,5,11] version 2.4.02 (http://strmix.esr.cri.nz).

Table 2
Experimental setup.

#	Contribs	Profiling kit	DNA of each contributor (pg)	Propositions: C = known contributor U = unknown RANDOM = simulated profile
1	1	GlobalFiler	400	Hp: RANDOM Hd: U
2	2	Profiler Plus	910:90	Hp: RANDOM + U Hd: 2U
3	2	GlobalFiler	200:200	Hp: RANDOM + U Hd: 2U
4	2	GlobalFiler	45:5	Hp: RANDOM + U Hd: 2U
5	3	GlobalFiler	100:67:33	Hp: RANDOM + 2U Hd: 3U

4. Results

We present results in the same format as given in [1], which we provide in Table 3. *LR*s generated in the comparisons of randomly simulated profiles to the evidence profiles were used to generate the distributions seen in Table 3. Also shown in Table 3 is the average H_d true *LR* value from all simulations, the *LR*s for the known contributors and p (the proportion of simulated non-donors that yield *LR*s greater than the known donors) expressed as '1 in' for each of the known contributor's *LR*s. p has been given as '1 in' for ease of direct comparison to the *LR* values of known contributors. We give this value because, as proven in [1], $1/p$ must be larger than or equal to the *LR* of the known contributor it is being compared against.

Due to the biassed manner in which genotypes are chosen, the distributions in Table 3 cannot simply be created by graphing a histogram of *LR*s obtained. Doing so would lead to an inflated count of the *LR* that corresponds to the genotype with the highest weight. Instead for each *LR* deciban (one tenth of a ban,$x \to x + 0.1$) the height, H, of the histogram bars (the count shown on the vertical axis in Table 3) are:

$$H_{x \to x+0.1} = \frac{I}{Y} \times \sum_{i: x < \log_{10}(LR_i) \leq x+0.1} b_i$$

for bracket $x \to x + 0.1$, where Y is the number of iterations (actual) for which the analysis was run and I is the extrapolated number of iterations a naïve simulator would have run to obtain an equivalent result. Due to the large value I, the height of the bars in histograms for commonly obtained *LR*s can hide other results. For example the distribution shown for simulation 4 in Table 3 when graphed on a log scale (as in Fig. 1) shows the counts of the distribution, otherwise hidden from view by scale.

The values for p (the proportion of non-donors who would yield a *LR* greater than or equal to that of the POI, LR_{POI}) can be recovered by:

$$p = \frac{I}{Y} \times \sum_{i: LR_i \geq LR_{POI}} b_i$$

5. Discussion

The number of simulations chosen in this demonstration (not the approximated naïve simulator number) was kept low deliberately to show the power of importance sampling in evaluating misleading *LR*s on highly discriminatory profiles. Table 3 shows that the number of simulations ranged from 1 to 10,000 even when discrimination power was over 25 orders of magnitude. The main driver in requiring greater numbers of simulations is the number of genotype sets resulting from the deconvolution that have moderately sized weights. This is expected to be largest for complex mixtures with roughly equally contributing donors.

The classic method of carrying out H_d true tests is to simulate profiles of non-donors for comparison to evidence DNA profiles. A limitation with this classic method of testing is that the profiles of the greatest importance (i.e. those which will give the highest *LR*s favouring their inclusion to the profile) are usually chosen a very

Table 3
Results of comparisons of simulated random references to profiles outlined in Table 1. Vertical axes for distributions show counts.

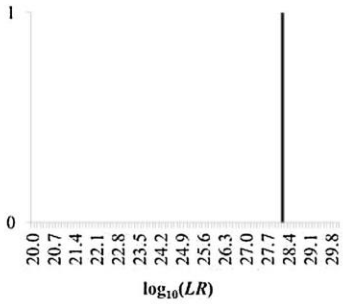
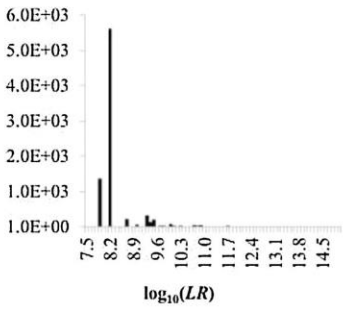
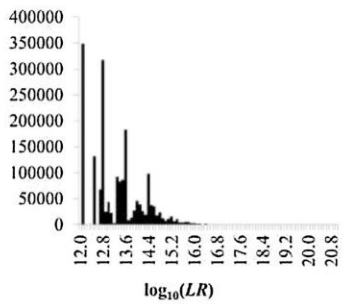
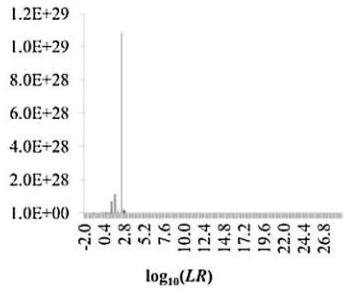
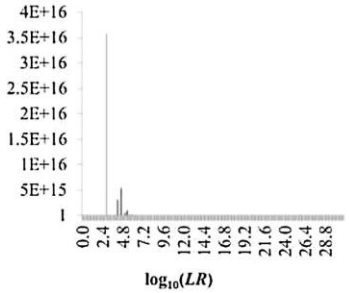
Experiment	Actual simulations (equivalent # naïve simulations)	H_d true $\log_{10}(LR)$ distribution	Average H_d true LR	Contributor number	H_p true LR (s)	p, '1 in'
1	1 (1.45×10^{28})		1.000	C_1	1.45×10^{28}	1.45×10^{28}
2	100 (3.21×10^{13})		0.97	C_1 C_2	5.29×10^{11} 1.06×10^{10}	9.62×10^{11} 1.46×10^{11}
3	1000 (1.12×10^{21})		1.12	C_1 C_2	6.54×10^{16} 1.22×10^{16}	4.34×10^{17} 9.35×10^{16}
4	10,000 (8.18×10^{31})		0.965	C_1 C_2	8.97×10^{27} 242	7.55×10^{28} 700

Table 3 (*Continued*)

Experiment	Actual simulations (equivalent # naïve simulations)	H_d true $\log_{10}(LR)$ distribution	Average H_d true LR	Contributor number	H_p true LR (s)	p, '1 in'
5	10,000 (7.24×10^{23})		0.998	C_1 C_2 C_3	2.40×10^{19} 2.73×10^{11} 1.03×10^{10}	1.05×10^{21} 3.10×10^{12} 1.84×10^{11}

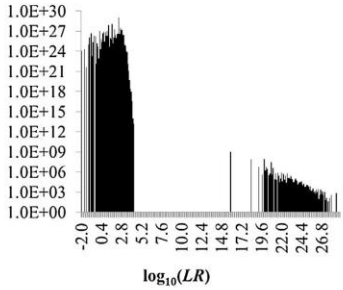


Fig. 1. LR distribution for simulation 4 when graphed on a log scale.

small proportion of the time. This makes H_d true tests of limited value for most modern DNA profiling kits, where *LR*s for known contributors can be 20 or more orders of magnitude. We have shown here how importance sampling can be used to overcome this problem, by biasing the choice of non-donor profiles and then adjusting for that bias in the evaluation of results. Using an importance sampling scheme it is still possible to recover all parameters of interest (i.e. the average *LR* or the proportion of non-donors that would yield a *LR* above a specific value). As the results in Table 3 show, the same diagnostics can still be interrogated with relatively low computation cost, and should hold true, specifically that the average H_d true *LR* is approximately 1 and that $p \leq LR_{POI}^{-1}$. Using importance sampling should provide the capability to carry out meaningful H_d true tests on almost all DNA profiles with relatively little computational cost.

Acknowledgements

This work was supported in part by grant 2011-DN-BX-K541 from the US National Institute of Justice. Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the U.S. Department of Justice. We acknowledge the comments of Catherine McGovern and two anonymous reviewers which have greatly improved this paper.

Appendix A. A simple example of importance sampling

We give, in this appendix, a simple example of importance sampling for readers not familiar with the technique. Our example is a problem involving a discrete probability distribution because the ideas are more easily demonstrated in the discrete case. Importance sampling works equally well with discrete distributions as it does with continuous distributions. The only difference is that the integral is replaced with a sum. That is, if $f(x)$, $p(x)$ and $h(x)$ are all discrete probability functions, then

$$E[f(X)] = \sum_{x \in \Omega_X} f(x) \frac{p(x)}{h(x)}$$

where Ω_X is the set of all possible outcomes for the random variable X (the sample space of X).

Imagine we are interested in estimating the probability of observing eight or more heads in ten tosses of a fair coin (fair $\equiv$ Pr(Heads) $= 0.5 =$ Pr(Tails)). We can calculate this probability directly with the Binomial distribution. If X is a random variable that represents the number of heads observed in ten tosses of a fair coin, then

$$\begin{aligned} \Pr(X \geq 8) &= \Pr(X = 8) + \Pr(X = 9) + \Pr(X = 10) \\ &= \binom{10}{8} 0.5^8 (1 - 0.5)^2 + \binom{10}{9} 0.5^9 (1 - 0.5)^1 \\ &\quad + \binom{10}{10} 0.5^{10} (1 - 0.5)^0 \\ &= \frac{56}{1024} = 0.0546875 \end{aligned}$$

If we wanted to calculate this value through simple Monte Carlo simulation, then we might take a sample of say $S = 10,000$ from a Binomial distribution with $n = 10$ and $p = 0.5$ ($X \sim Bin(n = 10, p = 0.5)$), and calculate

$$\Pr(X \geq 8) \approx \frac{1}{S} \sum_{i=1}^{S} I(x_i \geq 8)$$

where $I(x_i \geq 8)$ is an indicator function which takes the value of one when $x_i \geq 8$ and is zero otherwise. If we were to do this using

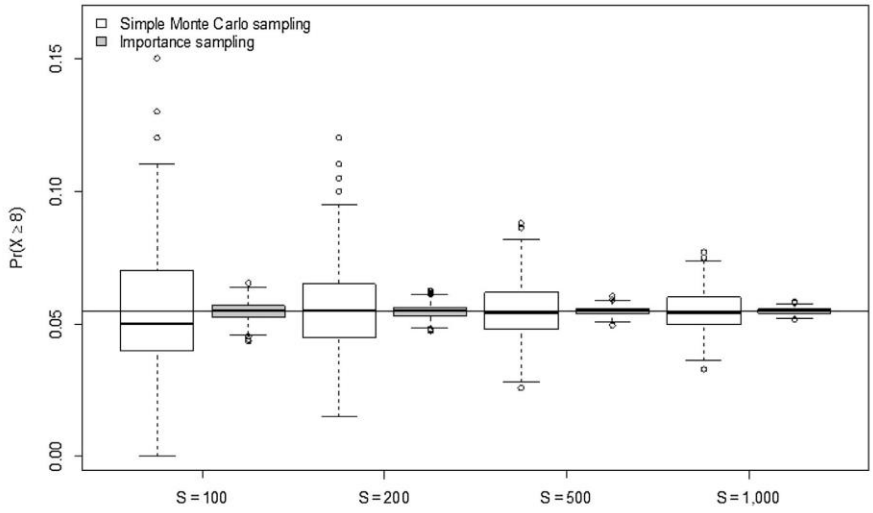


Fig. A2. S = 100–1000 experiments using importance sampling (grey) and naïve Monte Carlo (white) simulations to determine the probability of obtaining eight or more heads in 10 coin tosses. The theoretical probability, based on a binomial distribution, is shown as a solid line at 0.0547.

an importance sampling approach, then we might choose

$$h(x) = \begin{cases} \frac{1}{2} & ,x = 8 \\ \frac{1}{3} & ,x = 9 \\ \frac{1}{6} & ,x = 10 \end{cases}$$

as our importance distribution. This is simply a discrete distribution where the probabilities are in the same decreasing order (but not the same magnitude) as the binomial distribution on the values 8–10. Therefore, our importance weights in this sampling scheme are

$$w(x_i) = \begin{cases} \frac{\binom{10}{8}0.5^8(1-0.5)^2}{1/2} = \frac{90}{1024} & ,x_i = 8 \\ \frac{\binom{10}{9}0.5^9(1-0.5)^1}{1/3} = \frac{30}{1024} & ,x_i = 9 \\ \frac{\binom{10}{10}0.5^{10}(1-0.5)^0}{1/6} = \frac{6}{1024} & ,x_i = 10 \end{cases}$$

and our estimator is

$$\Pr(X \geq 8) \approx \frac{1}{S}\sum_{i=1}^{S} w(x_i) \quad .$$

In this example we can easily derive both the variance of the Monte Carlo estimator, and the variance of the importance sampling estimator. We can consider the Monte Carlo estimation as follows. We have a random variable X that follows a Binomial distribution with parameters $n = 10$ and $p = 0.5$. If we take a sample of size S, from this distribution and compute for each random variate $Z_i = I(X_i \geq 8)$, then Z_i follows a Bernoulli distribution with parameter $p = Pr(X \geq 8)$ which we denote p_8. Therefore, because the Z_i's are independent, the variance of the Monte Carlo estimator is

$$\begin{aligned} \mathrm{Var}\left[\frac{1}{S}\sum_{i=1}^{S} Z_i\right] &= \frac{1}{S^2}\sum_{i=1}^{S}\mathrm{Var}[Z_i] \\ &= \frac{1}{S^2}S p_8 (1-p_8) \\ &= \frac{p_8(1-p_8)}{S} \end{aligned}$$

If we think about the importance sampling estimator in the same way, then we generate a sample of independent random variates Y_i[2] with probability equal to the importance distribution $h(y)$. Therefore, we can firstly show that the importance sampling estimator is an unbiased estimator of the probability interest, because

$$\begin{aligned} \mathrm{E}\left[\frac{1}{S}\sum_{i=1}^{S} w(Y_i)\right] &= \frac{1}{S}\mathrm{E}\left[\sum_{i=1}^{S} w(Y_i)\right] \\ &= \frac{1}{S}S[w(8)\Pr(Y=8) + w(9)\Pr(Y=9) + w(10)\Pr(Y=10)] \\ &= \frac{\Pr(X=8)}{\Pr(Y=8)}\Pr(Y=8) + \frac{\Pr(X=9)}{\Pr(Y=9)}\Pr(Y=9) + \frac{\Pr(X=10)}{\Pr(Y=10)}\Pr(Y=10) \\ &= \Pr(X=8) + \Pr(X=9) + \Pr(X=10) = \Pr(X \geq 8) = p_8 \end{aligned}$$

The variance is derived in a similar manner:

$$\begin{aligned} \mathrm{Var}\left[\frac{1}{S}\sum_{i=1}^{S} w(Y_i)\right] &= \frac{1}{S^2}\mathrm{Var}\left[\sum_{i=1}^{S} w(Y_i)\right] \\ &= \frac{1}{S^2}S\sum_{y\in\{8,9,10\}} (w(y)-p_8)^2 h(y) \\ &= \frac{1}{S}\sum_{y\in\{8,9,10\}} (w(y)-p_8)^2 h(y) \end{aligned}$$

We can perhaps see, by inspection that the variance of the

[2] We use Y_i to recognize the difference in possible outcomes from the importance distribution $Y \in \{8, 9, 10\}$, as opposed to the original distribution $X \in \{0, 1, ..., 10\}$, therefore $h(y) = h(x)$.

importance sampling estimator is smaller than the variance of the Monte Carlo estimator. However it becomes immediately obvious when we evaluate the quantities numerically where we can show that the importance sampling estimator is approximately 42–47 times less variable than the Monte Carlo estimator. The extension of this is that the variance of the importance sampling estimator is zero when $w(y) = p_8$, i.e. if we sample from the conditional distribution of X given $X \geq 8$ then we would have a zero variance estimator for p_8. This is only possible if p_8 is known, and so has no practical use, but demonstrates that importance sampling is likely to work well when samples are drawn from a distribution that looks like the conditional distribution.

Therefore, we can regard the importance sample as being a sample of the weights which occur with probability defined by the probability distribution. We can also demonstrate this empirically. We repeated the two sampling Schemes 1000 times, using $S = 10,000$. This yields 1000 estimates of $\Pr(X \geq 8)$ from each method, which in turn allow us to see how well these sampling schemes estimate the quantity of interest, and how variable the estimates are. The ratio of the variances of each set of 1000 is 41.2 which is very close to our theoretical estimate.

Fig. A2 demonstrates the value of importance sampling, namely that (given a good importance distribution) it will often yield a more precise (less variable) estimate of the quantity of interest than that obtained by naïve Monte Carlo sampling for an equivalent sample size, or, alternatively it can give an estimate of similar precision for a much smaller sample size. In this example, a sample of size 250 yields estimates that are still less variable than naïve Monte Carlo estimates using a sample size of 10,000 (data not shown).

Appendix B. Derivation of Eq. (2)

We have in Eq. (1) that an approximation of the average LR (over the Y tests) that would have been obtained had a naïve simulator been used is:

$$\overline{LR} = \frac{1}{Y}\sum_{y} LR_y b_y$$

Had a naïve estimator been used (and assuming the importance sampler had been run for enough iterations that all the genotypes that lead to $LR > 0$ have been sampled) then all the iterations that the naïve simulator would have run, and have been skipped by the importance sampler, would have yielded $LR = 0$. We can then say that the sum of the LRs obtained from the importance sampler, plus all the zero LRs that would have been obtained (of which there would be $I - Y$ of them), divided by the total number of iterations that the naïve simulator would have run for (I) gives the average LR. In an equation this is:

$$\overline{LR} = \frac{(I - Y) \times 0 + \sum_{y} LR_y}{I}$$

Which, with some simplification and rearrangement gives Eq. (2) from the text:

$$I = \frac{\sum_{y} LR_y}{\overline{LR}}$$

References

[1] D. Taylor, J. Buckleton, I. Evett, Testing likelihood ratios produced from complex DNA profiles, Forensic Sci. Int.: Genet. 16 (2015) 165–171.
[2] M.D. Coble, J. Buckleton, J.M. Butler, T. Egeland, R. Fimmers, P. Gill, et al., DNA Commission of the International Society for Forensic Genetics: recommendations on the validation of software programs performing biostatistical calculations for forensic genetics applications, Forensic Sci. Int. Genet. 25 (2016) 191–197.
[3] Executive Office of the President: President's Council of Advisors on Science and Technology. Report to the president Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods (2016).
[4] D. Taylor, J.-A. Bright, J. Buckleton, The interpretation of single source and mixed DNA profiles, Forensic Sci. Int.: Genet. 7 (2013) 516–528.
[5] J.-A. Bright, D. Taylor, J.M. Curran, J.S. Buckleton, Developing allelic and stutter peak height models for a continuous method of DNA interpretation, Forensic Sci. Int.: Genet. 7 (2013) 296–304.
[6] I.J. Good, Probability and the Weighing of Evidence, Charles Griffin & Company Limited, London, 1950.
[7] H. Swaminathan, A. Garg, C. Grgicak, M. Medard, CEESIt: a computational tool for the interpretation of STR mixtures, Forensic Sci. Int.: Genet. 22 (2016) 149–160.
[8] K.-J. Slooten, T. Egeland, Exclusion probabilities and likelihood ratios with application to mixtures, Int. J. Legal Med. 130 (2016) 39–57.
[9] M. Kruijver, Efficient computations with the likelihood ratio distribution, Forensic Sci. Int.: Genet. 14 (2015) 116–124.
[10] D.A. Taylor, J.-A. Bright, J.S. Buckleton, The 'factor of two' issue in mixed DNA profiles, J. Theor. Biol. 363 (2014) 300–306.
[11] J.-A. Bright, D. Taylor, J.M. Curran, J.S. Buckleton, Degradation of forensic DNA profiles, Aust. J. Forensic Sci. 45 (2013) 445–449.

5b – Clarification
Further information on Appendix B

The effective sample size for importance sampling is the number of independent samples drawn from the target distribution to obtain an estimator with the same variance as the importance sampler.

A common method used to calculate ESS is given by:

$$ESS = \frac{\left(\sum_{i=1}^{Y} w_i\right)^2}{\sum_{i=1}^{Y} w_i^2}$$

Consider a simple example, where a simple evidence DNA profile, originates from a single contributor and the DNA profile of that contributor can be determined unambiguously (i.e. there is only 1 genotype that sensibly describes the evidence profile). This genotype as a population frequency of 1 in 1 million. A Monte Carlo simulation is run which randomly chooses DNA profiles from the population and compares them to the evidence profile in order to calculate an LR using propositions:

H1) The randomly drawn person is the source of the DNA

H2) Someone, unrelated to the randomly drawn individual, is the source of the DNA

We would expect that for every million iterations of the simulation we would obtain 999 999 profiles that were different to the evidence profile and so gave an LR = 0, and 1 iteration would sample a profile that matched the evidence profile and gave an LR of 1 million. Let the number of LRs of non-zero that were obtained be Y (in this case Y = 1) and the LR obtained for iteration y ($y \in Y$) be LR_y. Let the total number of iterations that the Monte Carlo simulation was run for be I ($Y \in I$). The number of zero LRs obtained is then ($I - Y$). The average LR can be calculated by:

$$\overline{LR} = \frac{(I-Y)\times 0 + \sum_{y} LR_y}{I}$$
$$= \frac{0+1000000}{1000000}$$
$$= 1$$

Now consider an importance sampling distribution of genotypes that contains only the genotype that will give a non-zero LR, therefore in this example containing 1 genotype. The weight associated with generating a sample, x, from this genotype from the proposal distribution, q(x), rather than the target distribution, $\pi(x)$, is:

$$w = \frac{q(x)}{\pi(x)} = \frac{\left(\frac{1}{LR}\right)}{1} = \frac{1}{LR}$$

Therefore, every iteration the importance sampler will choose the genotype that gives the non-zero *LR*, with associated weight LR^{-1}. In this instance there is zero variance in the values produced by the importance sampler and so matching of variances does not make sense to calculate ESS. If we use the formula above, then for *Y* samples from the proposed distribution:

$$ESS = \frac{\left(\sum_{i=1}^{Y} w_i\right)^2}{\sum_{i=1}^{Y} w_i^2} = \frac{\left(\frac{Y}{LR}\right)^2}{\left(\frac{Y}{LR^2}\right)} = Y$$

However, the analyst may wish to know how many iterations their importance sampling would be equivalent to, had they sampled directly from the target distribution. Now go back to the original derivation and consider the average *LR* obtained from the simulation using an importance sampling scheme were $\overline{LR}$. We consider that the importance sampler has run for enough iterations that all non-zero genotypes have been sampled from and so as well as the average *LR* we have *Y* (which is the number of importance samples taken) and the LRs at each iteration. The target is to obtain a value of I, so:

$$\overline{LR} = \frac{(I - Y) \times 0 + \sum_{y} LR_y}{I}$$

With rearrangement becomes:

$$I = \frac{\sum_{y} LR_y}{\overline{LR}} \approx \sum_{y} LR_y$$

With the approximation coming from the fact that given adequate draws, the average *LR* should be approximately 1. In the running example being used, we could consider *Y* importance samples so that:

$$I = \sum_{y} LR_y = Y(1000000)$$

i.e. each single importance sample yields an average *LR* that would be expected to take 1 million samples from the target distribution in order to achieve.

Manuscript: Do low template DNA profiles have useful quantitative data? D Taylor, J Buckleton. (2015) Forensic Science International: Genetics 16, 13-16 – *Cited 7 times*

Statement of novelty: This work carried out a comparison of semi-continuous and fully continuous DNA profile evaluation for very low level and complex DNA profile to demonstrate the fact that there is still information content in peak heights, even at low levels. Up until this point there had been no demonstration of this fact and opinions amongst the forensic community were divided.

My contribution: Main author and sole simulation programmer. Equal contributor to theory.

Research Design / Data Collection / Writing and Editing = 50% / 100% / 70%

Additional comments:

Forensic Science International: Genetics 16 (2015) 13–16

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

Do low template DNA profiles have useful quantitative data?

Duncan Taylor [a,b,*], John Buckleton [c]

[a] Forensic Science South Australia, 21 Divett Place, Adelaide, SA 5000, Australia
[b] School of Biological Sciences, Flinders University, GPO Box 2100, Adelaide, SA 5001, Australia
[c] ESR, Private Bag 92021, Auckland 1142, New Zealand

ARTICLE INFO

Article history:
Received 5 September 2014
Received in revised form 10 October 2014
Accepted 5 November 2014

Keywords:
DNA profile interpretation
Mixtures
Likelihood ratios
Low template
Continuous models

ABSTRACT

A set of low template mixed DNA profiles with known ground truths was examined using software that utilised peak heights (STRmix™ V2.3) and an adapted version that did not use peak heights and mimicked models based on a drop-out probability [1,2] (known as semi-continuous or 'drop' models) (STRmix™ lite). The use of peak heights increased the LR when H_p was true in the vast majority of cases. The effect was most notable at moderate template levels but was also present at quite low template levels. There is no level at which we can say that height information is totally uninformative. Even at the lowest levels the bulk of the data show some improvement from the inclusion of peak height information.

1. Introduction

Methods for evaluating DNA profiles have benefitted from recent improvements in modelling and software [2–8]. This has allowed the interpretation of many more profiles and the generation of a corresponding relevant match statistic, namely a likelihood ratio (LR) [3,7,9]. Previously Taylor [10] demonstrated that the LR generated by a continuous model [3] trended towards 1 as the template was reduced or as the number of contributors increased. This was true whether a true or false donor to the mixture was considered.

The question has been legitimately asked whether there is a point where the quantitative data present in peak height information becomes no more informative than merely the presence or absence of the peak.

There has been a view that profiles can be low template and that it is these low template profiles where peak height is of limited or no value. However for mixed DNA profiles it is likely that the contributors will be present in different template amounts. Hence each of, say, four contributors could be more or less low in template level. In addition nearly all casework profiles have a downward slope with respect to molecular weight. This is often termed a degradation slope. What this means is that it may be simplistic to refer to profiles as high or low template. Many profiles will exhibit a range of template estimates dependent on molecular weight. This observation has been made previously [11]. What this means is that the method used to interpret such profiles, which are prevalent in casework, must be able to interpret information that is very likely to range in template from high to low within the same profile or contributing component to a mixture.

There are three lemmas that can be considered useful at this point:

(1) Adding correct and relevant information to a calculation can only increase the ability to distinguish between a true and false proposition
(2) If the amount of information provided to a calculation is decreased, at some point the ability to distinguish between true and false propositions is entirely lost
(3) If a contributing profile can be determined unambiguously then additional information (such as peak height data) will not improve the ability of a calculation to distinguish between a true and false proposition

The hypothesis we wish to consider is: At low template the stochastic effects are such that the addition of peak height information to a calculation provides negligible additional ability to distinguish between true and false propositions.

Inline with lemma 3 there is little point trialling DNA profiles whose contributing genotypes can be determined unambiguously, as we know the addition of peak height information in these

* Corresponding author at: Forensic Science South Australia, 21 Divett Place, Adelaide, SA 5000, Australia. Tel.: +61 8 8226 7700; fax: +61 8 8226 7777.
E-mail address: Duncan.Taylor@sa.gov.au (D. Taylor).

http://dx.doi.org/10.1016/j.fsigen.2014.11.001
1872-4973/

instances will have no effect. What is needed are a set of low template mixed DNA profiles with ground truth known that have been analysed with and without the use of peak height information.

To this end we trial complex mixed DNA profiles with a range of input DNA, where the mixture proportions and donor profiles were known, to assess at what point peak height information no longer benefits an LR calculation, and specifically the ability of the LR to distinguish between known contributors (H_p true) and known non-contributors (H_d true).

2. Method

A range of four person mixtures produced in GlobalFiler (Thermo Fisher Scientific), as per manufacturer's instructions. Amplification fragments were resolved using the ABI PRISM® 3130xl Genetic Analyser and analysed in Genemapper® ID-X to obtain peak height information for each profile. These mixtures are samples 22–31 from [10], amplified in triplicate (note that there are only two replicates of sample 23 rather than three). We reproduce the relevant mixture information from [10] in Table 1.

As in [10] profiles were analysed down to 30 rfu.

STRmix™ version 2.3 (http://strmix.esr.cri.nz) was reconfigured to ignore peak height information. A probability of dropout, Pr(D), was required and was applied to each instance of a peak dropout. 1 − Pr(D) was applied to each instance of non-dropout. The value of Pr(D) was determined by including it as a parameter in the model, which sampled from its posterior with a flat prior by the Markov chain. This is effectively maximum likelihood estimation for Pr(D) for every profile during its analysis. Note that a single Pr(D) was applied across all contributors to the profile. This reconfigured STRmix™ product was termed STRmix™ lite and clones the performance of semi-continuous ('drop') models in a manner that is as close as possible to the normal functioning of STRmix. This was done, rather than using separate software, so that all factors other than the one of interest (peak height) would remain constant between the two experiments.

In both STRmix™ V2.3 and STRmix™ lite a uniform probability for allelic drop-in of 0.0017 was used inline with laboratory observations. To calculate LRs each combination of three individuals was assumed for each four person mixed DNA profile, meaning from the 29 profiles, 116 analyses were carried out and compared to a POI using the propositions:

H_p: The POI, contributor A, contributor B and contributor C are the sources of DNA

H_d: Contributor A, contributor B and contributor C and an unknown individual are the sources of DNA

where POI, A, B and C were combinations of contributors $\mathbf{C_{1,4}}$.

LRs were calculated using an in-house self-declared Caucasian GlobalFiler database and using the product rule.

Table 1
Mixture proportions and PCR setup.

Tubes	Mixture proportions for contributor				Total DNA added to PCR (pg)
	C_1	C_2	C_3	C_4	
22	0.25	0.25	0.25	0.25	400
23	0.25	0.25	0.25	0.25	200
24	0.25	0.25	0.25	0.25	50
25	0.25	0.25	0.25	0.25	20
26	0.25	0.25	0.25	0.25	10
27	0.40	0.30	0.20	0.10	400
28	0.40	0.30	0.20	0.10	200
29	0.40	0.30	0.20	0.10	50
30	0.40	0.30	0.20	0.10	20
31	0.40	0.30	0.20	0.10	10

3. Results

In Fig. 1 we give the improvement in the $\log_{10}$(LR) when peak height data is included in the analysis and plot against the input DNA of individual contributors. As input template level is not directly available from an electropherogram we also give the improvement against average allelic peak height of each profile.

Peaks below about 300 rfu are indicative of low template. Even profiles where the average peak height is 300 rfu often have low template components.

In Fig. 2 we plot the log of LR produced by STRmix™ lite vs the log of LR produced by STRmix™ V2.3. This allows an investigation of the benefit of peak heights for profiles that would produce a LR below 10^9 if peak height was not used.

In Table 2 we provide the results of H_d true testing (for an explanation of the H_d true test concept see Evett et al. [12] or Gill and Haned [13]) for sample 24, assuming C_2, C_3 and C_4 in both STRmix™ (left) and STRmix™ lite (right). H_d true tests are when non-donors are compared to a profile in order to generate an LR. In the propositions given in Section 2 we replace the POI with a DNA profile that has been randomly simulated in accordance with expectations from population allele frequencies. A large number of H_d true tests can be performed in order to give a series of 'diagnostics' about the profile analysis and in particular the performance of the models used to analyse the profile data. In Table 2 we provide the following values:

Simulations: The number of randomly simulated profiles that were compared to the evidence DNA profile

H_p true LR: The value of the LR obtained when compared with the known contributor

And for H_d true comparisons

p ('1 in'): p is the proportion of H_d true tests that yielded an LR at least as big as the LR obtained from the known contributor. Values give are the inverse of p so that they can be directly compared to the size of the H_p true LR

LR = 0: The percentage of simulations that resulted in an LR of zero being obtained

LR ≥ 1: The percentage of simulations that resulted in an LR that favoured the inclusion of the randomly simulated non-donor. These have classically considered as 'false inclusions'.

Average LR: The average value of all the H_d true LRs. Theory predicts that this value should be one as explained in Good [14] (quoting Turing).

4. Discussion

The results presented in Fig. 2 demonstrate a strong advantage in using peak height information down to very low levels. We see that the $\log_{10}$(LR)s with and without peak height information converge towards 0 (LR = 1) as the information in the profile diminishes. Hence both approaches are correctly reporting that the profile becomes uninformative at extremely low template.

Most (110/116) instances of including peak height information yielded a higher LR when compared to known contributors. The greatest effect can be seen for results at high LRs. These differences are of less practical importance.

However for LRs less than one billion in STRmix™ lite (i.e. to the left of the vertical line in Fig. 2) there is still significant increase in LRs when peak height information is included. In many instances this is four or five orders of magnitude. There is no level at which we can say that height information is totally uninformative. Even

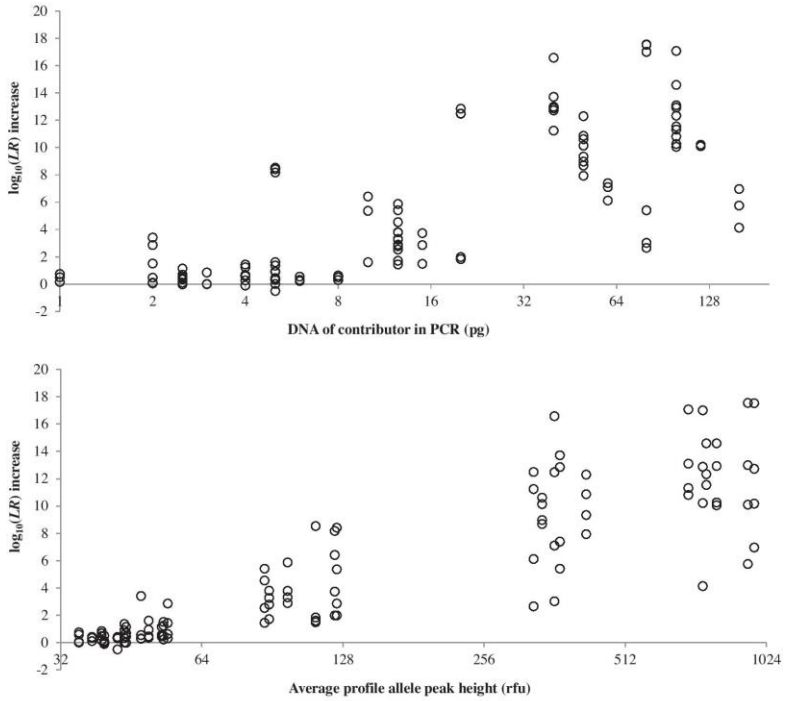


Fig. 1. The $\log_{10}$(LR) increase obtained by including peak height information against input DNA (top) or average profile peak height (bottom).

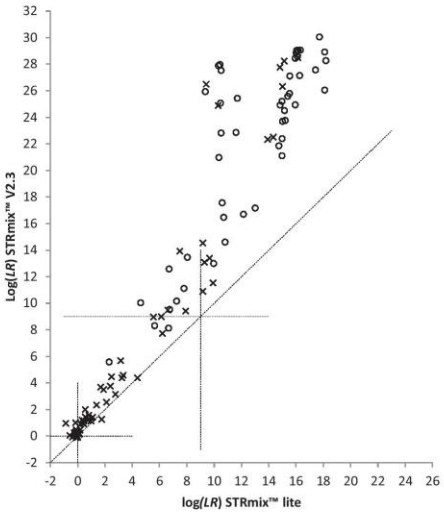


Fig. 2. $\log_{10}$(LR) for STRmix™ lite and STRmix™ V2.3. The diagonal dashed line is $x = y$. Circles represent values that were derived from greater than 50 pg of total input DNA, and crosses less than 50 pg.

Table 2
Results of H_d true tests for a four person 0.25:0.25:0.25:0.25 mix at 50 pg total input template. Average peak height for the profile was 89 rfu.

	STRmix™ V2.3	STRmix™ lite
# Simulations	12,000,000	10,000,000
H_p true LR	374,104	207
H_d true		
p ('1 in')	3,000,000	11,947
LR = 0	99.958%	94.491%
LR > 1	0.0173%	0.0472%
Average LR	1.005	1.078

at the lowest levels the bulk of the data are above the diagonal line. From Fig. 2 the majority of data points indicate that for an LR of 10^x generated without using peak height information there is between a $10^{x/2}$ and 10^x fold increase in LR when peak height information is added.

Also vital is to consider what effect the inclusion of peak height data may have on comparisons to known non-contributors, i.e. when H_d is true. Inspection of Table 2 indicated markedly increased H_d true LRs above 1, if peak height is ignored. Also seen in Table 2, when ignoring peak height information, are increased H_d true LRs above the H_p true LR (p in Table 2). These could be termed false inclusions or adventitious matches and should be minimised.

We therefore conclude from this work that at low template DNA levels the inclusion of peak height information can have a

very significant and beneficial impact on the LR produced from comparisons to both known contributor and non-donors. In the DNA profiles and LRs considered within this work a beneficial impact was seen in most instances.

Acknowledgements

This work was supported in part by grant 2011-DN-BX-K541 from the US National Institute of Justice. Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the U.S. Department of Justice. We gratefully acknowledge comments of Catherine McGovern and Stuart Cooper which have greatly improved this paper.

References

[1] P. Gill, J.P. Whitaker, C. Flaxman, N. Brown, J.S. Buckleton, An investigation of the rigor of interpretation rules for STR's derived from less that 100 pg of DNA, Forensic Sci. Int. 112 (2000) 17–40.
[2] D.J. Balding, J. Buckleton, Interpreting low template DNA profiles, Forensic Sci. Int. Genet. 4 (2009) 1–10.
[3] D. Taylor, J.-A. Bright, J. Buckleton, The interpretation of single source and mixed DNA profiles, Forensic Sci. Int. Genet. 7 (2013) 516–528.
[4] H. Haned, Forensim, An open-source initiative for the evaluation of statistical methods in forensic genetics, Forensic Sci. Int. Genet. 5 (2011) 265–268.
[5] K. Lohmueller, N. Rudin, Calculating the weight of evidence in low-template forensic DNA casework, J. Forensic Sci. 58 (2013) 234–259.
[6] C.D. Steele, D.J. Balding, Statistical evaluation of forensic DNA profile evidence, Annu. Rev. Stat. Appl. 1 (20) (2014) 1–4.
[7] R. Puch-Solis, L. Rodgers, A. Mazumder, S. Pope, I. Evett, J. Curran, et al., Evaluating forensic DNA profiles using peak heights, allowing for multiple donors, allelic dropout and stutters, Forensic Sci. Int. Genet. 7 (2013) 555–563.
[8] M.W. Perlin, M.M. Legler, C.E. Spencer, J.L. Smith, W.P. Allan, J.L. Belrose, et al., Validating TrueAllele® DNA mixture interpretation, J. Forensic Sci. 56 (2011) 1430–1447.
[9] R.G. Cowell, S.L. Lauritzen, J. Mortera, Probabilistic modelling for DNA mixture analysis, Forensic Sci. Int. Genet. 1 (2008) 640–642.
[10] D. Taylor, Using continuous DNA interpretation methods to revisit likelihood ratio behaviour, Forensic Sci. Int. Genet. 11 (2014) 144–153.
[11] P. Gill, J. Buckleton, A universal strategy to interpret DNA profiles that does not require a definition of low-copy-number, Forensic Sci. Int. Genet. 4 (2010) 221–227.
[12] I.W. Evett, J.A. Lambert, J.S. Buckleton, B.S. Weir, Statistical analysis of a large file of STR profiles of British Caucasians to support forensic casework, Int. J. Legal Med. 109 (1996) 173–177.
[13] P. Gill, H. Haned, A new methodological framework to interpret complex DNA profiles using likelihood ratios, Forensic Sci. Int. Genet. 7 (2013) 251–263.
[14] I.J. Good, Probability and the Weighing of Evidence, Charles Griffin & Company Limited, London, 1950.

Chapter 6: Placing the theoretical model into practise

By 2015, STRmix™ had been in use in most forensic labs around Australia and New Zealand for one to three years. The forensic community decided to revisit the question that sparked all this work off in the first place, i.e. are different laboratories and analysts getting consistent LRs from the same evidence. The last time this had been attempted was during the 'crisis' meeting of 2009 at which point it was clear that forensic biology laboratories had far to go to reach a state of equal justice outcomes.

A study was set up whereby a series of profiles were sent to forensic laboratories around Australia and New Zealand and they were asked to analyse and interpret the profiles (comparing reference profiles that were provided as part of the study) and report the LR as they would for a normal criminal case. The setup of this study and the findings make up the publication in this section.

The study found a dramatic improvement in the consistency between labs compared to the 2009 attempt. The biggest source of variation was now the choice of number of contributors. This then sparked another arm of work that is discussed in the next chapter.

Another source of variation that was identified from the study was whether the use of peak(s), below the analytical (or detection) threshold should be considered when analysts pre-assessed the DNA profile to determine the number of contributors. The use of STRmix™ generally caused laboratories to drop their analytical threshold so that the additional, low level, peaks could be used in evaluations. From this practise, it would have been expected that the presence of these sub-threshold peaks to be more of an issue in the pre-STRmix™ days. On review, it was found that the presence of sub-threshold peaks were indeed more prevalent pre-STRmix™, however, did not have much impact because those profiles that contained data such as this were almost always deemed unsuitable for interpretation (in fact the presence of sub threshold peaks tended to be one of the decision point in making this determination). With STRmix™ providing a means to evaluate so many more DNA profiles than before, the issue was brought to the forefront. One option was to continue to consider DNA profiles that possessed sub-threshold peaks as not suitable for interpretation, however conceptually it doesn't sit well that a strong and well resolved profile could be evaluated, but the presence of a small 'blip' in the baseline could render it uninterpretable. The disconnect causing the issue existed between the information the analyst was using to interpret the profile and the information being provided to STRmix™. The solution developed to address the disconnect was to incorporate a way for users to provide prior beliefs in the mixture proportions to STRmix™. This way, if a user saw sub threshold peaks that indicated a very low-level contributor may be present, a prior mixture proportion could be supplied that indicated a contributor had most of its mass at low levels (near 0). This spawned an arm of work that lead to the second publication in this section, which explores the prevalence, impact and solutions to sub-threshold data.

Manuscript: Investigating a common approach to DNA profile interpretation using probabilistic software. S Cooper, C McGovern, JA Bright, D Taylor, J Buckleton. (2015) Forensic Science International: Genetics 16, 121-131 – *Cited 3 times*

Statement of novelty: This work compares the consistency of laboratories using STRmix™ around Australian and New Zealand. At the time of publication such a comparison of laboratory performance had not previously been done and published in Australia or New Zealand.

My contribution: Minor role as author and theorist. Forensic Science SA was one of the participant laboratories in the study and I was the coordinator for the SA participants i.e. explanation of study, dissemination of material and compilation of results.

Research Design / Data Collection / Writing and Editing = 10% / 0% / 10%

Additional comments:

Forensic Science International: Genetics 16 (2015) 121–131

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

Investigating a common approach to DNA profile interpretation using probabilistic software

Stuart Cooper [a,*], Catherine McGovern [a], Jo-Anne Bright [a], Duncan Taylor [b,c], John Buckleton [a]

[a] ESR, Private Bag 92021, Auckland 1142, New Zealand
[b] Forensic Science South Australia, 21 Divett Place, Adelaide, SA 5000, Australia
[c] School of Biological Sciences, Flinders University, GPO Box 2100, Adelaide, SA, Australia

ARTICLE INFO

Article history:
Received 2 October 2014
Received in revised form 18 December 2014
Accepted 23 December 2014

Keywords:
Forensic DNA interpretation
STRmix™
Mixed DNA profile
Standardisation

ABSTRACT

Recently there has been a drive for standardisation of DNA profile interpretation within and between different forensic laboratories. The continuous interpretation software STRmix™ has been adopted for use by laboratories in Australia and New Zealand for profile interpretation. Within this paper we examine the concordance in profile interpretation of three crime samples by twenty different analysts across twelve different international laboratories using STRmix™. The three profiles selected for this study exhibited a range of template and complexity. The use of probabilistic software has compelled a level of concordance between different analysts however there remain differences within profile interpretation, particularly with the objective assignment of the number of contributors to profiles.

1. Introduction

The interpretation of forensic DNA profiles is complicated by allelic dropout and drop-in, artefacts such as stutter and the presence of DNA from more than one individual, termed mixed DNA profiles [1]. There is significant variation across forensic laboratories in the methods used for profile interpretation. In a recent study of 107 laboratories (including US federal, state and local and three international laboratories) approximately 30% indicated that they use the cumulative probability of inclusion (CPI), 56% used the random match probability (RMP) [2] and 14% used a likelihood ratio (*LR*). This variation was also observed within the 12 different participating laboratories/institutes within this study, where prior to the availability of continuous statistical methods, approximately 17% indicated they may have used a CPI/random man not excluded (RMNE) method, depending on the results, 41.5% used RMP and 41.5% used *LR*.

The *LR* has very significant advantages for the interpretation of ambiguous profiles [3]. Ambiguous profiles include mixtures and profiles where allelic dropout and drop-in are likely. The definition of an *LR* is the probability of the DNA profiling evidence (E) given two competing propositions. One typically aligns with the prosecution point of view, H_p, and the other the defence, H_d, given all the relevant information, I:

$$LR = \frac{\Pr(E|H_p, I)}{\Pr(E|H_d, I)}$$

Traditional methods of interpretation are often described as binary, so called because the probability of the evidence given a proposed genotype is assigned either as zero or one (see for example [4]). Methods have been developed that use the ratio of the peak heights (or peak areas) of the two alleles at a heterozygote locus (heterozygous balance or peak height ratio), and mixture proportions when considering mixed DNA profiles, to determine whether combinations of genotypes were supported or not [5,6]. In 2012, the editors of the Journal of Forensic Science International: Genetics highlighted the need for more research in the area of DNA profile interpretation and the creation of statistical software packages to advance development and facilitate implementation of the generally accepted standards in forensic genetics [7].

Known shortcomings of the binary model [8,9] have led to the development of probabilistic models that factor in the probability of dropout and drop-in [10–14]. Semi-continuous methods do not explicitly use peak heights when generating possible genotype sets and do not explicitly model artefacts such as stutter.

Recently, a number of fully continuous probabilistic methods have been described that model allelic and/or stutter peaks within a DNA profile [15–19]. Continuous methods evaluate the probability of a set of peak heights given proposed genotype sets. These

* Corresponding author. Tel.: +64 9 845 1726; fax: +64 9 849 6046.
E-mail address: Stuart.cooper@esr.cri.nz (S. Cooper).

http://dx.doi.org/10.1016/j.fsigen.2014.12.009
1872-4973/

methods are designed to be used in expert systems to remove much of the requirement for the subjective assignment of peaks as alleles or stutter within evidence profiles. A further discussion of the merits of the different interpretation models can be found in Kelly et al. [20] and Steele and Balding [21].

A number of studies have investigated variability in DNA profile interpretation both within a single laboratory and across jurisdictional laboratories [22–25]. The authors highlight a high degree of diversity in the interpretation methods used for mixed DNA profiles. Furthermore, the interpretation was shown to be highly subjective. In particular the findings of Dror and Hampikian [22] reported in New Scientist under the title "Fallible DNA evidence can mean prison or freedom" [23] have been used to highlight the subjectivity in DNA evidence interpretation and some would argue the bias inherent to certain methodologies. Recently there has been a drive for standardisation in DNA profile interpretation across different jurisdictions [26–29].

In 2012 STRmix™, an interpretation software that employs a continuous model [17,30,31], was introduced into the Australian and New Zealand forensic DNA laboratories for profile interpretation. STRmix™ can interpret a wide variety of DNA profiles, from single source DNA rich samples to complex, low level mixtures with multiple contributors. The model uses the quantitative information from an electropherogram (epg), such as peak height data and molecular weight, to calculate the probability of the peak heights given all possible genotype combinations, assigning a weight to each possible genotype set. The weights are determined using Markov chain Monte Carlo (MCMC).

Continuous interpretation methods remove some thresholds, such as heterozygote balance and stutter ratio (*SR*), and thus help remove some elements of the subjective decision making required within a binary model. Following the interpretation of the DNA profiling data, STRmix™ can provide a statistical assessment in the form of an *LR* from the comparison of a person of interest's reference DNA profile.

STRmix™ currently requires the analyst to assign the number of contributors to the profile prior to analysis. An uncertain number of contributors can be a source of variability in the *LR* [32,33]. The MCMC is another source of variability [34]. The implementation of the same software across different laboratories has been shown to reduce but not necessarily eliminate the variability between analysts analysing the same DNA profiling results [35]. The present study was undertaken to assess the level of standardisation in profile interpretation achieved by the introduction of the continuous DNA interpretation model STRmix™. Within this paper we investigate whether standardisation could be achieved by implementing the same probabilistic software within and between different laboratories. It is worth highlighting that while standardisation may well help us to achieve 'the same' answer this should not be at the expense of achieving the 'right answer'.[1]

To evaluate whether there was a difference in inter and intra laboratory variability, nine members of staff from the one laboratory and eleven participants from international jurisdictional laboratories or consultancies were invited to undertake independent analysis of a number of mixed DNA profiles. We describe the results of this collaborative study within this paper.

2. Methods

Twenty different participants were asked to undertake the interpretation of DNA profiles from three casework scenarios. Nine members of staff from one laboratory undertook independent analysis in addition to eleven international participants. This cohort included participants from Australia, New Zealand, the US, Canada and the UK. The intra-laboratory participants were assigned a random number between 1 and 9 with the remaining participants assigned a number between 10 and 20. All participants had previously undertaken a STRmix™ training course. We note that this may homogenise opinions more than a random sample of analysts.

The exercise involved the assessment of Identifiler™ DNA profiles generated from three questioned samples. As the samples were casework samples, the true number of contributors to each sample was unknown. Relevant case circumstances, STRmix™ input files capturing both allelic and stutter peak height data and reference DNA profiles were provided for comparison and generation of an *LR* as required. All questioned sample epgs (including zooms of the baseline) were supplied to participants. Further details are provided in Appendix A. A summary of the reference DNA profiles supplied for each case is provided in Appendix B.

Samples were profiled using the Applied Biosystems' Identifiler™ (Life Technologies, CA) PCR amplification kit at 28 cycles and separated by capillary electrophoresis using a 3130*xl* Genetic Analyzer following the manufacturer's recommended settings. Peak height data were captured using GeneMapper™ v3.2.1 applying a 50 rfu analytical threshold with no stutter filter applied. Participants were asked to comment on whether, and if so by what means, they would have provided any form of statistical analysis in relation to these three case samples prior to implementation of STRmix™ using the following scale:

a Too complex, no interpretation progressed.
b Potentially include or exclude a given reference, however no statistic provided.
c Calculate a statistic.
d Other, please comment.

Participants were asked to review the questioned sample epgs, determine the likely number of contributors and suitable propositions given the information provided and undertake an interpretation using STRmix™. Where appropriate *LR*s were calculated assigning appropriate prosecution and defence propositions (H_p,H_d).

To restrict additional variation the STRmix™ parameters were provided to the participants. Specific details of the STRmix™ settings are available upon request from the authors. These settings were informed by empirical data relevant to the specific laboratory protocols used to generate the DNA profiling results. This included a zero drop-in rate for the standard profiling analysis method employed. The participants were asked to use a Caucasian allele frequency database which was also provided [36].

Participants were asked to undertake the analyses as they saw fit and submit their findings once their interpretations were complete. As a means of assessing each analysis the point estimate *LR*s (i.e. with no sampling uncertainty undertaken) were selected and these are displayed in Figs. 1–3. Submissions summarised by the number of contributors assigned the propositions used to undertake an *LR* and the point estimate *LR*s are provided in Appendix C.

3. Results

Table 1 is a summary of the responses to the question: How would you have dealt with the findings prior to the use of STRmix™? A tally of the different institutes responses ($n = 12$) is given with comments below. We have considered participants

[1] One can debate whether there is a 'right answer'. Please read this as the answer most supported by current knowledge.

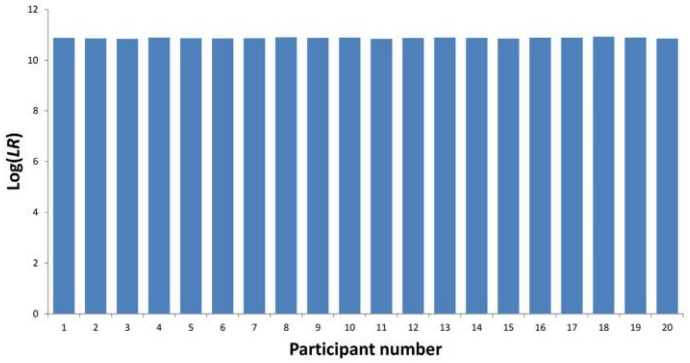


Fig. 1. Log (*LR*) calculated by each participant for Sample 1.

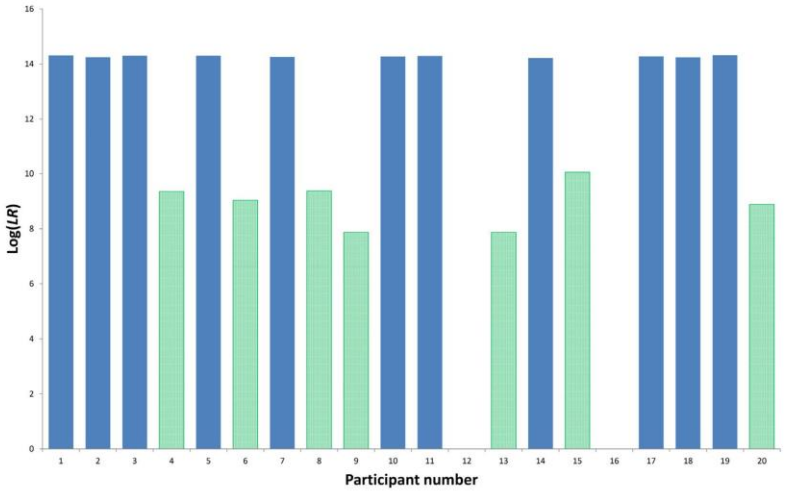


Fig. 2. Log (*LR*) for each participant, where calculated, for Sample 2. The solid bars are log (*LR*s) considering two contributors and the lighter/patterned bars are considering three contributors.

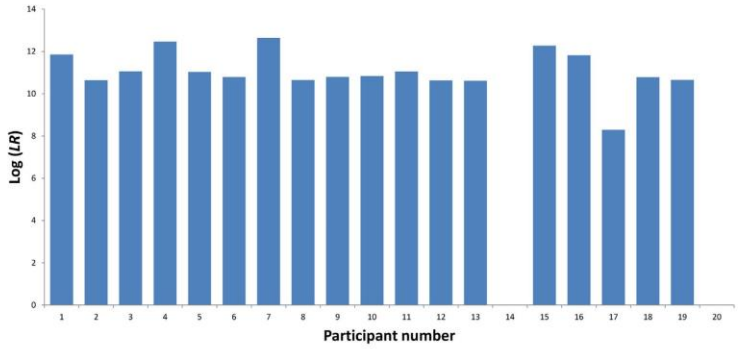


Fig. 3. Log (*LR*), where calculated, by each participant for Sample 3.

Table 1
Summary of the different institutes responses ($n = 12$) to how would they have dealt with these cases prior to having access to STRmix™.

Sample	Too complex, no interpretation	Include or exclude references, no statistic given	Calculate a statistic[a]	Other
1	–	–	12	–
2	2[b,c]	4[d]	6	–
3	8	3[d]	1	–

[a] The statistical method that would have been employed varied and included random man not excluded (RNME), random match probability (RMP), and likelihood ratio (*LR*).
[b] A response from the laboratory with 9 participants indicated they may have gone on to include or exclude, but likely no statistic would have been provided given the low peak heights.
[c] Participant 13 stated they would consider the major only and that the minor was unsuitable. They may step up to include or exclude for the major only. In this instance, as the major contributor did not correspond with the person of interest (POI), then no statistic would have been undertaken. A statistic would have been undertaken if POI corresponded with major.
[d] Participant 12 stated they would likely just include or exclude, and give no statistic. However, could have provided a statistic if required.

1–9 as providing one joint response here. We consider the responses for each case in turn.

3.1. Sample 1

All participants were confident to assign this profile as a two person mixture, providing an *LR* using the following propositions:

H_p: The DNA originated from the complainant and an unknown individual.

H_d: The DNA originated from two unknown individuals.

A summary of the *LR*s calculated for each analyst is given in Fig. 1.

In addition, most participants ($n = 16$) compared the profile to the person of interest (POI) using the following propositions:

H_p: The DNA originated from the POI and an unknown individual.

H_d: The DNA originated from two unknown individuals.

All comparisons resulted in an *LR* of 0; an exclusion. The remaining four participants stated they would have excluded the POI 'by eye' based on the assumption of two contributors to the mixture, so did not calculate an *LR* in STRmix™. A table of all the *LR*s provided by the participants can be found in Appendix C.

3.2. Sample 2

The majority of participants ($n = 18$) undertook at least one STRmix™ analysis and provided an *LR* for a comparison to the POI reference sample (refer Fig. 2). Participants 12 and 16 did not undertake a STRmix™ analysis of this mixture. Participants 12 and 16 stated they would re-amplify this sample in order to clarify the number of contributors. Eleven participants assumed two contributors and the remaining seven assumed three contributors to the profile. All participants who calculated an *LR* used the following propositions:

H_p: The DNA originated from the POI and one (or two) unknown individuals.

H_d: The DNA originated from two (or three) unknown individuals.

Six of the participants carried out secondary STRmix™ analyses, such as increasing or decreasing the contributor number by one, i.e. if they had assigned the profile as a two person mixture, they also provided an *LR* for a three person mixture. A table of all the *LR*s provided by the participants can be found in Appendix C.

3.3. Sample 3

Eighteen participants calculated an *LR* for this profile assuming the presence of three contributors using the following propositions (refer Fig. 3):

H_p: The DNA originated from the POI and two unknown individuals.

H_d: The DNA originated from three unknown individuals.

Participants 14 and 20 did not undertake interpretation of this mixture. Participant 14 stated they were not able to assign a likely number of contributors. Participant 20 stated that this appeared to be at least a four person mixture and therefore was not suitable for analysis within their laboratory. Participant 17 used the results of only one of the two replicates provided in order to carry out the analysis following their laboratory protocol. This has resulted in a reduction of the profile information available to the interpretation and a lower comparative *LR* as a result.

Two of the participants who provided an *LR* for a three person mixture also undertook an additional analysis and *LR* calculation assuming four contributors. A table of all the *LR*s provided by the participants can be found in Appendix C.

4. Conclusions

Sample 1 appears to demonstrate an example of a good quality two person mixture, where the amount of DNA present from each contributor is in approximately equal proportions. All participants assumed the same number of contributors and the *LR*s provided have a high degree of reproducibility (average log (LR) = 10.36, SD = 0.02). As expected the *LR*s were not identical due to the Monte Carlo aspect of the MCMC within STRmix™. This is a relatively new source of variability within the forensic DNA analysis process, although this has been highlighted by other groups [19,37]. Importantly the *LR*s are very similar and the variability is very small compared to other sources of variability in forensic DNA profiling including PCR as demonstrated by Bright et al. [38].

If we compare this highly reproducible set of *LR*s in Fig. 1 to what would have theoretically been provided from the range of the statistical methods used by participants prior to implementation of STRmix™ (refer Table 1) then STRmix™ implementation can be viewed as a positive step towards standardisation. To illustrate, Participant 15 undertook a RMP calculation for Sample 1 in relation to the complainant, using an in-house program and provided a value of 1 in 2.23×10^9 (calculated per NRC II recommendation 4.1, $\theta = 0.01$, using a Caucasian allele frequency database), an order of magnitude smaller than the 20 match statistics provided in this study. Although the practical consequence of a difference of one order of magnitude given the relatively high *LR* is negligible. The results of Sample 1 indicate that in certain circumstances it is possible to achieve a high level of concordance both within and between jurisdictional laboratories for the interpretation and statistical assessment of the same DNA evidence.

An element of variation was observed in Sample 1 results when it came to the consideration of the POI reference profile. This DNA profile, whilst having all alleles represented within the mixed DNA profile, could be readily excluded as a plausible contributor given an assumption of two contributors. At this point two approaches to interpretation were observed: either to progress an *LR* calculation ($LR = 0$) to underpin a conclusion of exclusion, or to make an exclusionary statement without a supporting statistic. Whilst we

could debate the theoretical purity versus pragmatism of these two approaches, this is unlikely to translate into any discernible difference to the strength of evidence provided to the court.

Further to the *LR*s submitted some participants provided comments in relation to Sample 1. For example some discussed that under the assumption of two contributors the POI was excluded, however, the POI did share a significant number of alleles in common with a true donor to this mixture and so, from an intelligence point of view a relative of the POI may be a plausible consideration for investigators. One participant suggested further analysis, such as Y-STR profiling, to explore the number of male contributors and potentially address whether a male relative of the POI could be a consideration. From an exploratory point of view two participants hypothesised increasing the number of contributors by one to facilitate some additional form of statistical evaluation in relation to the POI. However, both these participants and the authors of this paper advocate that an assessment of the number of contributors be made in advance of sighting any reference DNA samples that cannot be assumed as legitimate contributors to the evidence, and where appropriate this assumption is not affected by the subsequent provision of reference DNA profiling information. A change in interpretation strategy in order to obtain a favourable pre-determined outcome (i.e. an inclusionary *LR*) must be avoided for a fair assessment of the evidence. This concept of bias and sequential unmasking is discussed in [22,39]. We would however, support a change in the assigned number of contributors if, after initial interpretation, a review of the output, including mixture proportions and genotype combination weights, indicated that the original assumption was incorrect. For transparency, we would promote that all such interpretations remain on file for defence disclosure. In addition, the receipt of additional relevant information, such as the provision of a reference profile from a consenting partner in a sexual assault case, could also change our assigned number of contributors.

Sample 2 represents a mixture of DNA with an increased level of ambiguity due to many components of the profile being detected below the stochastic threshold, and with some indication of additional information below the analytical threshold. These results raised the issue of assigning an appropriate estimate of the number of contributors to DNA profiling results. Overall most participants were confident to assign a number of contributors and undertake an *LR*. However, two participants were not.

It is worth highlighting that regardless of the number of contributors assigned where an *LR* was calculated, all exceeded one million. Hence all would fall at the extremely strong scientific support level if using the verbal scale of support described by Evett and Weir [40].

When interpreted as a two person mixture the *LR*s have a higher degree of reproducibility (average log (LR) = 14.28, SD = 0.03) than when interpreted as a three person mixture (average log (LR) = 8.92, SD = 0.74).

Many participants acknowledged that the profile contained a suggestion of the presence of a third contributor, predominantly due to peaks observed in the epg below the analytical threshold. In addition to the *LR*s, a number of participants also provided comments about this case, such as: Is 50 rfu a suitable analytical threshold? Could more information be obtained by decreasing the threshold? In this case the answer may well be yes. This analytical threshold (AT) was the implemented threshold for the laboratory who produced the crime profiles, and AT values are often set high to be conservative. Lowering the AT is a risk assessment, balancing the potential to yield more data with the increased risk for inclusion of artefacts such as raised baseline, pull-up and capillary carryover [41] in the profile.

Three participants undertook an exploratory mixture deconvolution assuming a three person mixture. However, after a review of the mixture proportions and genotype weightings assigned by STRmix™ the participants questioned if these were what they had anticipated when compared to their initial expectations based on their review of the epg. When analysed as a two person mixture the proportions and weights are more intuitive i.e. better reflected what was observed in the epg. The comments suggested the expectation prior to any STRmix™ analysis for this profile, assuming it was a three person mixture, would be for a major:minor:trace type profile. Generally speaking for those who ran this as a three person mixture the mixture proportions suggested by STRmix™ were closer to 1:1:1. This is a consequence of having little or no data available for STRmix™ to use for the proposed trace third contributor and forcing STRmix™ to spread the weights between the available genotypes. The use of the STRmix™ outputs to interrogate an analyst's initial assessment of the DNA profiling data is strongly recommended and may help inform or prompt further testing as discussed next.

A number of participants proposed that a re-amplification of this sample may have been beneficial to potentially obtain further information about the minor component, and to better inform the estimation of the number of contributors.

Assigning the number of contributors for Sample 2 is a point of difference both within the same laboratory and between laboratories. With the introduction of continuous models this has become one of the most discussed aspects of mixture interpretation and may be one area which requires further insight. Nevertheless, the introduction of a continuous interpretation model has facilitated a change in the type of sample from which an *LR* can be provided. For this profile only six participants stated they would have undertaken some form of statistical analysis prior to STRmix™. This increased to 18 participants calculating some form of LR statistic when provided with STRmix™.

Sample 3 represented a complex mixture of DNA where replicate profiling results were provided. All participants who undertook analysis (n = 18) and provided an *LR* did so assuming this was a three person mixture. The results illustrate an increased degree of variation in the *LR*s when compared to Samples 1 and 2, which is likely due to the complex nature of the results (average log (LR) = 11.21, SD = 0.68 [omitting participant 17]).

Again, a number of participants provided comments in relation to this profile. It became apparent from the feedback provided that analysts value replicate amplification of complex DNA profiling results such as Sample 3 because additional profiling information increases their confidence and informs their estimate of the number of contributors. As STRmix™ has the ability to consider replicate profiles from the same DNA extract this is something we advocate, wherever appropriate.

Interestingly, both participants who did not provide an *LR* for Sample 3 stated that using their previous interpretation methods they may have reported that the POI could not be excluded; however, no statistic would (or could) have been provided.

Overall the results from participants 1 through 9 demonstrated that in certain circumstances a level of standardisation can be achieved within one laboratory. All participants from within the one laboratory assumed three contributors and provided *LR*s (average log (LR) = 11.32, SD = 0.74). As for Sample 2, assigning a number of contributors again became the key point of difference between laboratories.

Sample 3 also demonstrated that a continuous interpretation model can increase the number of samples for which an interpretation and statistical evaluation can be progressed from 1 participant indicating they would provide a statistic prior to STRmix™ to 18 participants post STRmix™ access.

5. Discussion

This study demonstrates that for mixed DNA profiling results where the number of contributors is not ambiguous it is possible to achieve a standardised, consistent approach to the interpretation and statistical assessment of DNA evidence. This concept helps give confidence to the criminal justice system that irrespective of where and by whom a DNA profiling result is analysed a consistent statistical assessment of the evidence can be achieved.

Both the study and on-going dialogue within the forensic community continue to highlight that the confident estimation and assignment of the number of contributors to DNA profiling results are essential to our ability to effectively interpret DNA profiles. This topic has been debated for some time and alternative methods have been put forward to attempt to address such issues [42,43]. We suggest that the 'true' number of contributors to any questioned sample is always unknown. What we, as experienced analysts can provide is a meaningful estimation of the number of contributors which is well supported by the data and based on our knowledge of DNA profile behaviour in order to facilitate an interpretation. STRmix™ is a tool and as such may be used by an analyst for exploratory examination of the number of contributors and propositions. The subsequent choice of statistic to report depends on laboratory policy and we have seen different approaches including reporting the lowest or average *LR*. Where possible, we would suggest limiting the number of different statistics reported within a statement to ease comprehension. We do agree however, that some profiles are simply too complex and may never be able to be assigned a number of contributors with any degree of certainty.

With the implementation of the same probabilistic software there remains a level of subjectivity notably if there is a requirement to assign a likely number of contributors. This is an area which we advocate the need for further investigation. However, a greatly improved degree of standardisation can be achieved by implementation of the same probabilistic software within and between laboratories.

Acknowledgements

The input and expertise of all the participants is gratefully appreciated. These were (in no particular order):

Zane Kerr, New South Wales Forensic & Analytical Science Service, Australia,

Dr. Michael Coble, National Institute of Standards & Technology, USA,

Kimberley Windram, Forensic Science South Australia,

Dr. Timothy Kalafut, US Army Criminal Investigation Laboratory & Dr. Brenda Held, US Army Defence Forensic Science Centre,

Lisa Federle, Victoria Police Forensic Services Centre, Australia,

Dr. Ian Evett & Dr Sue Pope, Principal Forensic Services, UK,

Carl Grosser, Forensic Science Service Tasmania, Australia,

Steven Myers, California Department of Justice, USA,

Emma Caunt, Forensic & Scientific Services, Department of Health, Queensland, Australia,

Linda Parker & Dr. Jonathan Millman, Centre of Forensic Sciences, Toronto, Ontario, Canada,

Julie Murakami, Forensic Biology, Pathwest, Western Australia,

Richard Wivell, ESR, New Zealand,

Sue Vintiner, ESR, New Zealand,

Johanna Veth, ESR, New Zealand,

Pauline Simon, ESR, New Zealand,

Lisa Melia, ESR, New Zealand,

Kate Stevenson, ESR, New Zealand,

Turlough Thomas-Stone, ESR, New Zealand,

Maryanne Kregting, ESR, New Zealand,

Timothy Power, ESR, New Zealand.

This work was supported in part by grant 2011-DN-BX-K541 from the US National Institute of Justice. Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the U.S. Department of Justice.

Appendix A. Case 1 (Sample 1)

Case 1 (Sample 1)

A knife was located, discarded in an alleyway in relation, to an alleged assault. This was a separate and unrelated location to that of the incident. However, it is believed to have been used during the incident.

The knife was submitted to the laboratory for analysis. A sample of the bloodstain located on the item had been submitted for DNA analysis and had produced the mixed DNA profile attached.

Reference DNA samples were submitted from the complainant and a person of interest (POI) in this matter.

The investigators wish to know – could any DNA present have originated from the complainant?

Furthermore could any of the DNA present have originated from the POI?

Case 2 (Sample 2)

A glove was located at the scene of a burglary. This does not belong to the home owners and they have had no contact with the item.

A sample from the inside of the glove had been submitted for DNA analysis and had produced the DNA profile attached. A reference DNA sample was submitted from a POI who denies having any contact with the glove.

Investigators wish to know whether any DNA present on the glove could have originated from the POI.

Case 3 (Sample 3)

A glass pipe seized in relation to a drugs related incident was submitted to the laboratory for analysis.

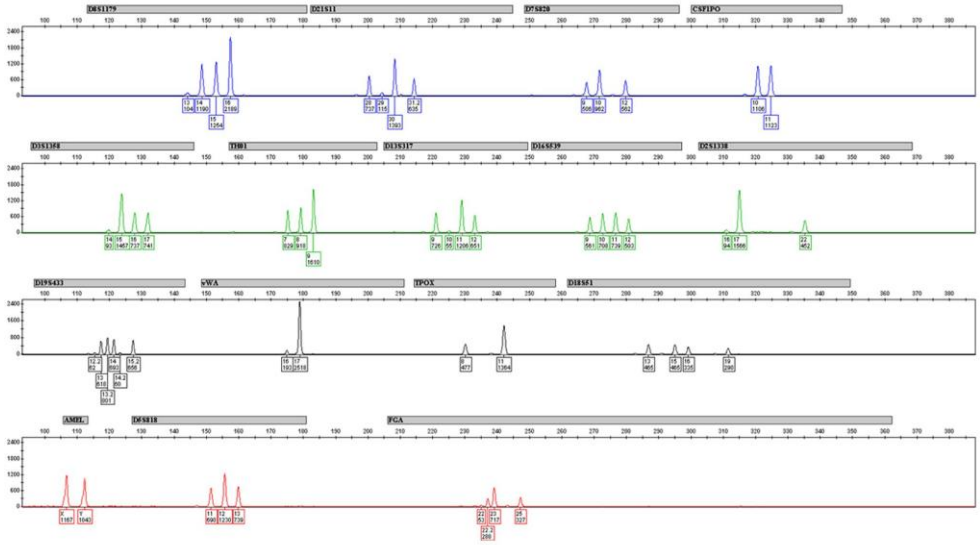
A sample taken from the mouthpiece of the pipe had undergone DNA profiling analysis and had produced the mixed DNA profile attached.

As the profile was low level and complex, it was amplified twice.

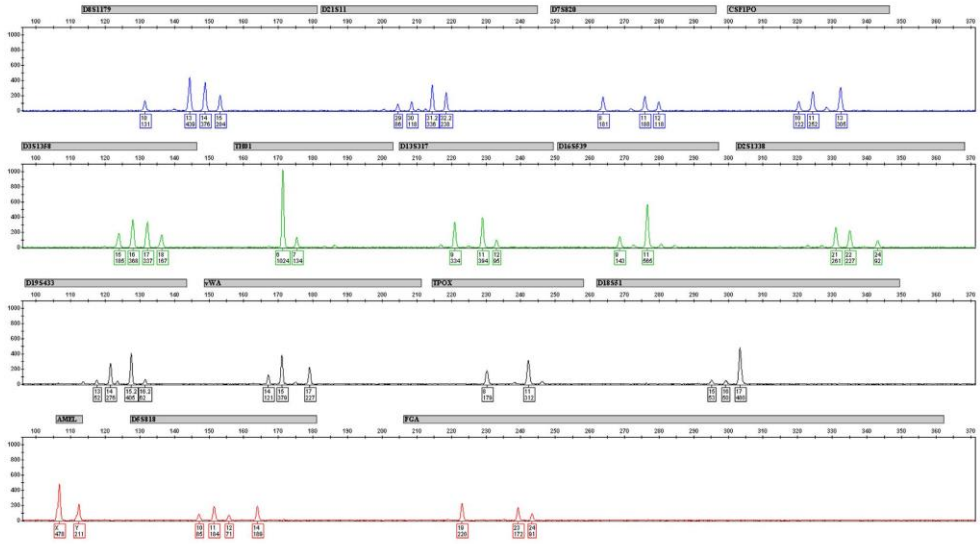
A POI has come to light and a reference DNA sample has also been provided for comparison.

Investigators wish to know whether any DNA present on the pipe could have originated from the POI.

Case 1 – Questioned Sample 1

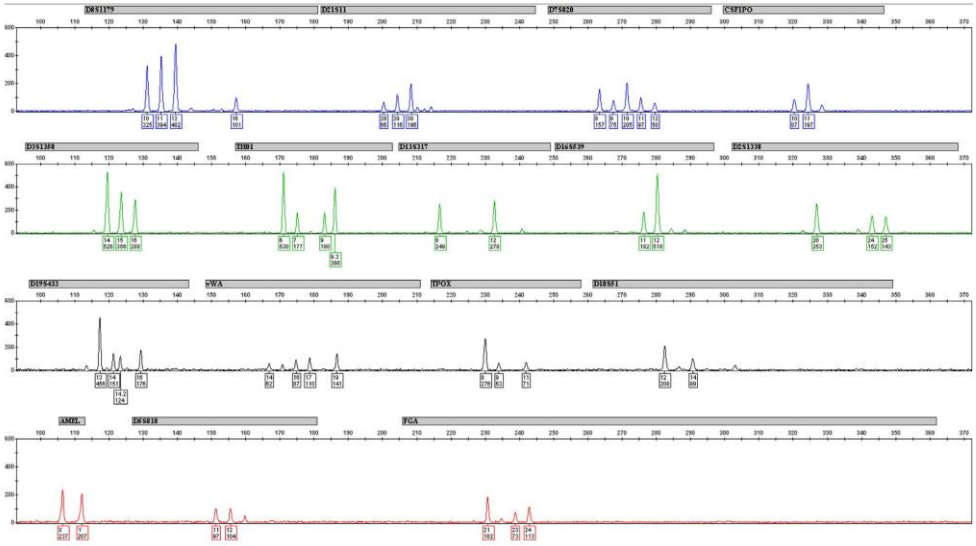


Case 2 – Questioned Sample 2

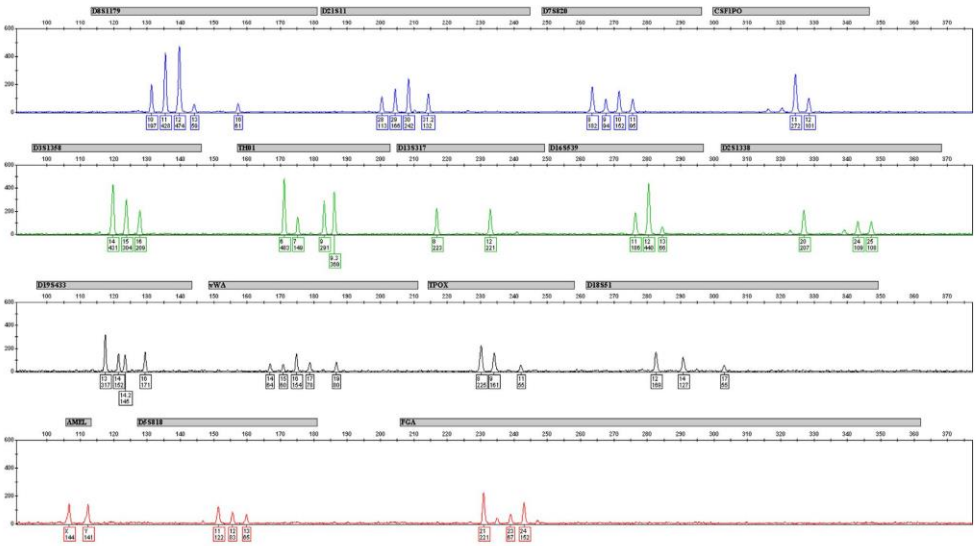


Case 3 – Questioned Sample 3

Amplification 1 (replicate 1):



Amplification 2 (replicate 2):



Appendix B.

Summary of the reference DNA profiles provided for Cases 1 through 3

Case 1

	D8	D21	D7	CSF	D3	TH01	D13	D16	D2	D19	vWA	TPOX	D18	Amel	D5	FGA
Complainant	14, 15	28, 30	10, 12	10, 11	15, 17	7, 9	11, 11	10, 11	17, 22	13,13.2	17, 17	8, 11	13, 19	X, Y	12, 13	22.2, 25
POI	14, 14	28, 28	12, 12	10, 10	16, 16	7, 8	9, 11	9, 10	22, 22	13, 14	17, 17	8, 11	13,15	X, Y	11, 11	23, 25

Case 2

	D8	D21	D7	CSF	D3	TH01	D13	D16	D2	D19	vWA	TPOX	D18	Amel	D5	FGA
POI	10, 15	29, 30	11, 12	10, 13	15, 18	6, 7	11, 12	9, 11	21, 24	15.2, 16.2	14, 15	11, 12	15, 17	X, X	10, 12	19, 24

Case 3

	D8	D21	D7	CSF	D3	TH01	D13	D16	D2	D19	vWA	TPOX	D18	Amel	D5	FGA
POI	10, 12	29, 30	8, 10	11, 11	14, 14	6, 9.3	8, 12	12, 12	20, 25	13, 16	16, 19	8, 9	12, 14	X, Y	11, 12	21, 24

Appendix C.

Case 1: Summary of LRs provided by participants (n = *20*)

Where POI, person of interest; U, unknown; NA, not applicable.

Participant number	Number of contributors assigned	Propositions H_p/H_d	LR (point estimate)	Propositions H_p/H_d	LR (point estimate)
1	2	C + U/U + U	7.65*E* + 10	NA	Not calculated. Excluded if two person mixture
2	2	C + U/U + U	7.34*E* + 10	POI + U/U + U	0
3	2	C + U/U + U	7.07*E* + 10	POI + U/U + U	0
4	2	C + U/U + U	7.78*E* + 10	POI + U/U + U	0
5	2	C + U/U + U	7.40*E* + 10	POI + U/U + U	0
6	2	C + U/U + U	7.29*E* + 10	POI + U/U + U	0
7	2	C + U/U + U	7.40*E* + 10	POI + U/U + U	0
8	2	C + U/U + U	8.14*E* + 10	POI + U/U + U	0
9	2	C + U/U + U	7.62*E* + 10	NA	Not calculated. Excluded if two person mixture
10	2	C + U/U + U	7.80*E* + 10	POI + U/U + U	0
11	2	C + U/U + U	7.06*E* + 10	POI + U/U + U	0
12	2	C + U/U + U	7.58*E* + 10	POI + U/U + U	0
13	2	C + U/U + U	7.83*E* + 10	POI + U/U + U	0
14	2	C + U/U + U	7.69*E* + 10	POI + U/U + U	0
15	2	C + U/U + U	7.20*E* + 10	POI + U/U + U	0
16	2	C + U/U + U	7.73*E* + 10	POI + U/U + U	0
17	2	C + U/U + U	7.71*E* + 10	NA	Not calculated. Excluded if two person mixture
18	2	C + U/U + U	8.49*E* + 10 10 K/50 K burnin/accepts	NA	Not calculated. Excluded if two person mixture
19	2	C + U/U + U	7.95*E* + 10	POI + U/U + U	0
20	2	C + U/U + U	7.21*E* + 10	POI + U/U + U	0

Case 2: Summary of LRs provided by participants (n = 20)

Where POI, person of interest; U, unknown; NA, not applicable.

Participant number	Number of contributors assigned	Propositions H_p/H_d	LR (point estimate)	Additional analysis	LR (point estimate)
1	2	POI + U/U + U	2.03*E* + 14		
2	2	POI + U/U + U	1.77*E* + 14		
3	2	POI + U/U + U	1.99*E* + 14		
4	3	POI + U + U/ U + U + U	2.25*E* + 09		
5	2	POI + U/U + U	1.99*E* + 14	Also analysed as a 3 person mixture: POI + U + U/ U + U + U	4.19*E* + 08
6	3	POI + U + U/ U + U + U	1.09*E* + 09	Also analysed as a 4 person mixture: POI + U + U + U/ U + U + U + U	6.24*E* + 08
7	2	POI + U/U + U	1.81*E* + 14	Also analysed as a 3 person mixture: POI + U + U/ U + U + U	1.27*E* + 11
8	3	POI + U + U/ U + U + U	2.36*E* + 09		
9	3	POI + U + U/ U + U + U	7.46*E* + 07		
10	2	POI + U/U + U	1.88*E* + 14	Also analysed as a 3 person mixture: POI + U + U/ U + U + U	9.67*E* + 08
11	2	POI + U/U + U	1.96*E* + 14	Also analysed as a 3 person mixture: POI + U + U/ U + U + U	3.82*E* + 08
12	Unable to determine	NA	NA		
13	3	POI + U + U/ U + U + U	7.43*E* + 07		
14	2	POI + U/U + U	1.65*E* + 14		
15	3	POI + U + U/ U + U + U	1.12*E* + 10	Also analysed as a 2 person mixture: POI + U/U + U	1.96*E* + 14
16	Inconclusive	NA	NA		
17	2	POI + U/U + U	1.90*E* + 14		
18	2	POI + U/U + U	1.74*E* + 1410 K/50 K burnin/ accepts		
19	2	POI + U/U + U	2.09*E* + 14	Also analysed omitting D18: POI + U/U+ U	4.57*E* + 13
20	3	POI + U + U/ U + U + U	7.69*E* + 08		

Case 3

Summary of LRs provided by participants ($n = 20$).
Where POI, person of interest; U, unknown.

Participant number	Number of contributors proposed	Propositions H_p/H_d	LR (point estimate)	Additional analysis	LR (point estimate)
1	3	POI + U + U/U + U + U	7.14*E* + 11		
2	3	POI + U + U/U + U + U	4.33*E* + 10		
3	3	POI + U + U/U + U + U	1.14*E* + 11		
4	3	POI + U + U/U + U + U	2.93*E* + 12		
5	3	POI + U + U/U + U + U	1.06*E* + 11		
6	3	POI + U + U/U + U + U	6.16*E* + 10		
7	3	POI + U + U/U + U + U	4.37*E* + 12		
8	3	POI + U + U/U + U + U	4.49*E* + 10		
9	3	POI + U + U/U + U + U	6.31*E* + 10		
10	3	POI + U + U/U + U + U	6.80*E* + 10		
11	3	POI + U + U/U + U + U	1.13E + 11	Also analysed as a 4 person mixture: POI + U + U + U/U + U + U + U	1.04*E* + 10
12	3	POI + U + U/U + U + U	4.22*E* + 10		
13	3	POI + U + U/U + U + U	4.04*E* + 10		
14	Unable to determine	NA	NA		
15	3	POI + U + U/U + U + U	1.88*E* + 12	Also analysed as a 4 person mixture: POI + U + U + U/U + U + U + U	9.70*E* + 09
16	3	POI + U + U/U + U + U	6.66*E* + 11		
17	3	POI + U + U/U + U + U	1.95*E* + 08		
18	3	POI + U + U/U + U + U	6.04*E* + 10		
19	3	POI + U + U/U + U + U	4.51*E* + 10		
20	Likely 4	NA	NA		

References

[1] P. Gill, L. Gusmão, H. Haned, W.R. Mayr, N. Morling, W. Parson, et al., DNA commission of the International Society of Forensic Genetics: recommendations on the evaluation of STR typing results that may include drop-out and/or drop-in using probabilistic methods, Forensic Sci. Int. Genet. 6 (2012) 679–688.

[2] T. Bille, J.-A. Bright, J. Buckleton, Application of random match probability calculations to mixed STR profiles, J. Forensic Sci. 58 (2) (2013) 474–485.

[3] P. Gill, C.H. Brenner, J.S. Buckleton, A. Carracedo, M. Krawczak, W.R. Mayr, et al., DNA commission of the International Society of Forensic Genetics: recommendations on the interpretation of mixtures, Forensic Sci. Int. 160 (2006) 90–101.

[4] T.M. Clayton, J.S. Buckleton, Mixtures, in Forensic DNA Evidence Interpretation, CRC Press, Boca Raton, 2004, pp. 217–274.

[5] T. Clayton, J.P. Whitaker, R.L. Sparkes, P. Gill, Analysis and interpretation of mixed forensic stains using DNA STR profiling, Forensic Sci. Int. 91 (1998) 55–70.

[6] P. Gill, R.L. Sparkes, R. Pinchin, T. Clayton, J.P. Whitaker, J.S. Buckleton, Interpreting simple STR mixtures using allelic peak areas, Forensic Sci. Int. 91 (1998) 41–53.

[7] A. Carracedo, P.M. Schneider, J. Butler, M. Prinz, Focus issue—analysis and biostatistical interpretation of complex and low template DNA samples, Forensic Sci. Int. Genet. 6 (6) (2012) 677–678.

[8] J. Buckleton, C.M. Triggs, Is the 2p rule always conservative? Forensic Sci. Int. 159 (2006) 206–209.

[9] J.S. Buckleton, C.M. Triggs, S.J. Walsh, Forensic DNA Evidence Interpretation, CRC Press, Boca Raton, Florida, 2004.

[10] H. Haned, Forensim: an open-source initiative for the evaluation of statistical methods in forensic genetics, Forensic Sci. Int. Genet. 5 (4) (2011) 265–268.

[11] H. Haned, P. Gill, Analysis of complex DNA mixtures using the forensim package, Forensic Sci. Int. Genet. 3 (1) (2011) e79–e80.

[12] D.J. Balding, J. Buckleton, Interpreting low template DNA profiles, Forensic Sci. Int. Genet. 4 (1) (2009) 1–10.

[13] A.A. Mitchell, J. Tamariz, K. O'Connell, N. Ducasse, Z. Budimlija, M. Prinz, et al., Validation of a DNA mixture statistics tool incorporating allelic drop-out and drop-in, Forensic Sci. Int. Genet. 6 (6) (2012) 749–761.

[14] K. Lohmueller, N. Rudin, Calculating the weight of evidence in low-template forensic DNA casework, J. Forensic Sci. 58 (1) (2013) 234–259.

[15] R.G. Cowell, S.L. Lauritzen, J. Mortera, A gamma model for DNA mixture analyses, Bayesian Anal. 2 (2) (2007) 333–348.

[16] I.W. Evett, P.D. Gill, J.A. Lambert, Taking account of peak areas when interpreting mixed DNA profiles, J. Forensic Sci. 43 (1) (1998) 62–69.

[17] D. Taylor, J.-A. Bright, J. Buckleton, The interpretation of single source and mixed DNA profiles, Forensic Sci. Int. Genet. 7 (5) (2013) 516–528.

[18] R. Puch-Solis, L. Rodgers, A. Mazumder, S. Pope, I. Evett, J. Curran, et al., Evaluating forensic DNA profiles using peak heights, allowing for multiple donors, allelic dropout and stutters, Forensic Sci. Int. Genet. 7 (5) (2013) 555–563.

[19] M.W. Perlin, M.M. Legler, C.E. Spencer, J.L. Smith, W.P. Allan, J.L. Belrose, et al., Validating TrueAllele® DNA mixture interpretation, J. Forensic Sci. 56 (2011) 1430–1447.

[20] H. Kelly, J.-A. Bright, J.S. Buckleton, J.M. Curran, A comparison of statistical models for the analysis of complex forensic DNA profiles, Sci. Justice 54 (1) (2014) 66–70.

[21] C.D. Steele, D.J. Balding, Statistical evaluation of forensic DNA profile evidence, Annu. Rev. Stat. Appl. 1 (2014) 20–21.

[22] G. Dror, G. Hampikian, Subjectivity and bias in forensic DNA mixture interpretation, Sci. Justice 51 (4) (2011) 204–208.

[23] L. Geddes, Fallible DNA evidence can mean prison or freedom, New Sci. 207 (2773) (2010) 8–11.

[24] J.M. Butler, Mixture interpretation: lessons from interlab study MIX05, National CODIS Conference, Arlington, VA, 2006.

[25] M.D. Coble, MIX13: an interlaboratory study on the present state of DNA mixture interpretation in the U.S., 5th Annual Prescription for Criminal Justice Forensics, Fordham University School of Law, 2014.

[26] Scientific working group on DNA analysis methods (SWGDAM). SWGDAM Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Testing Laboratories. [available from: http://www.fbi.gov/hq/lab/html/codis_swgdam.pdf] (2010).

[27] N. Morling, I. Bastisch, P. Gill, P.M. Schneider, Interpretation of DNA mixtures–European consensus on principles, Forensic Sci. Int. Genet. 1 (3–4) (2007) 291–292.

[28] Victoria lifts ban on DNA evidence. The Courier Mail 12 January 2010 [available from: http://www.couriermail.com.au/news/victoria-lifts-ban-on-dna-evidence/story-e6freon6-1225818464644?nk=2320bdc12d729f7cb718305a0b0408f3].

[29] A. Haelser, Issues in gathering, interpreting and delivering DNA evidence, Expert Evidence Conference, Canberra, 2011.

[30] J.-A. Bright, D. Taylor, J.M. Curran, J.S. Buckleton, Developing allelic and stutter peak height models for a continuous method of DNA interpretation, Forensic Sci. Int. Genet. 7 (2) (2013) 296–304.

[31] J.-A. Bright, D. Taylor, J.S. Buckleton, Degradation of forensic DNA profiles, Aust. J. Forensic Sci. 45 (4) (2013) 445–449.

[32] J.-A. Bright, J.M. Curran, J.S. Buckleton, The effect of the uncertainty in the number of contributors to mixed DNA profiles on profile interpretation, Forensic Sci. Int. Genet. 12 (2014) 208–214.

[33] J.-A. Bright, D. Taylor, J. Curran, J. Buckleton, Searching mixed DNA profiles directly against profile databases, Forensic Sci. Int. Genet. 9 (2014) 102–110.

[34] D. Taylor, J.A. Bright, J. Buckleton, J. Curran, An illustration of the effect of various sources of uncertainty on DNA likelihood ratio calculations, Forensic Sci. Int. Genet. 11 (2014) 56–63.

[35] L. Prieto, H. Haned, A. Mosquera, M. Crespillo, M. Alemañ, M. Aler, et al., Euroforgen-NoE collaborative exercise on LRmix to demonstrate standardization of the interpretation of complex DNA profiles, Forensic Sci. Int. Genet. 9 (2014) 47–54.

[36] J.-A. Bright, J.S. Buckleton, C.E. McGovern, Allele frequencies for the four major sub-populations of New Zealand for the 15 identifiler loci, Forensic Sci. Int. Genet. 4 (2) (2010) e65–e66.

[37] J. Curran, A MCMC method for resolving two person mixtures, Sci. Justice 48 (4) (2008) 168–177.

[38] J.-A. Bright, K.E. Stevenson, J.M. Curran, J.S. Buckleton, The variability in likelihood ratios due to different mechanisms, Forensic Sci. Int. Genet. 14 (0) (2015) 187–190.

[39] D.E. Krane, S. Ford, J.R. Gilder, K. Inman, A. Jamieson, R. Koppl, et al., Sequential unmasking: a means of minimizing observer effects in forensic DNA interpretation, J. Forensic Sci. 53 (4) (2008) 1006–1007.

[40] I.W. Evett, B.S. Weir, Interpreting DNA Evidence–Statistical Genetics for Forensic Scientists, Sinauer Associates, Inc., Sunderland, 1998.

[41] Life Technologies Corporation. Considerations for Evaluating Carryover on Applied Biosystems Capillary Electrophoresis Platforms in a HID Laboratory. https://tools.lifetechnologies.com/downloads/tech-doc-carryover-ce-systems-during-hid-analysis.pdf (2012).

[42] D. Taylor, J.A. Bright, J. Buckleton, Interpreting forensic DNA profiling evidence without specifying the number of contributors, Forensic Sci. Int. Genet. 13 (2014) 269–280.

[43] R.G. Cowell, T. Graversen, S. Lauritzen, J. Mortera, Analysis of forensic DNA mixtures with artefacts. arXiv:1302.4404 (2013).

Manuscript: Does the use of probabilistic genotyping change the way we should view sub-threshold data? D Taylor, J Buckleton, JA Bright. (2017) Australian Journal of Forensic Sciences 49 (1), 78-92 – *Cited 1 time*

Statement of novelty: This work addresses the issues surrounding the use of a threshold in a continuous DNA interpretation system. The exploration of the topic itself came from practical issues being faced by those using continuous DNA interpretation systems. A potential statistical solution is provided to the presence of sub-threshold information by supplying user-informed prior beliefs into the analysis.

My contribution: Main author and sole simulation programmer. Equal contributor to theory.

Research Design / Data Collection / Writing and Editing = 33% / 100% / 50%

Additional comments:

Australian Journal of Forensic Sciences, 2015
http://dx.doi.org/10.1080/00450618.2015.1122082

Taylor & Francis
Taylor & Francis Group

Does the use of probabilistic genotyping change the way we should view sub-threshold data?

Duncan Taylor[a,b]*, John Buckleton[c,d] and Jo-Anne Bright[c]

[a]*Biology Department, Forensic Science South Australia, Adelaide, SA, Australia;* [b]*School of Biological Sciences, Flinders University, Adelaide SA, Australia;* [c]*ESR, Auckland, New Zealand;* [d]*National Institute of Standards and Technology, Gaithersburg, MD, United States of America*

(*Received 11 August 2015; accepted 15 November 2015*)

The sensitivity and resolution of modern DNA profiling hardware is such that forensic laboratories generate more data than they have resources to analyse. One coping mechanism is to set a threshold, above the minimum required by instrument noise, so that weak peaks are screened out. In binary interpretations of forensic profiles, the impact of this threshold (sometimes called an analytical threshold – AT) was minimal as interpretations were often limited to a clear major component. With the introduction of continuous typing systems, the interpretation of weak minor components of mixed DNA profiles is possible and consequently the consideration of peaks just above or just below the analytical threshold becomes relevant. We investigate here the occurrence of low-level DNA profile information, specifically that which falls below the analytical threshold. We investigate how it can be dealt with and the consequences of each choice in the framework of continuous DNA profile interpretation systems. Where appropriate we illustrate how these can be implemented using the probabilistic interpretation software STRmix. We demonstrate a feature of STRmix that allows the analyst to guide the software, using human observation that there is a low-level contributor present, through user-designated prior distributions for contributor mixture proportions.

Keywords: DNA profile interpretation; sub-threshold; likelihood ratios; analytical threshold

1. Introduction

The primary method for the analysis of a DNA sample is amplification by polymerase chain reaction (PCR), which incorporates a flurophore. This is followed by separation of the fragments by capillary electrophoresis. The output is a trace of fluorescence versus time that is referred to as an electropherogram (epg). Most laboratories set an analytical threshold (AT), above which peaks are labelled at analysis. The AT is often set well above the level of electronic noise. Peaks in the epg may be artefactual or allelic. Epg analysis software can recognise and filter some of the well-characterised artefacts, but many still require the judgement of a human analyst. Many of these remaining artefactual peaks can be recognised by position or morphology. In binary interpretations, the impact of these weak peaks was minimal as interpretations were often limited to the interpretation of a clear major component. With the introduction of continuous typing systems the interpretation of weak minor components of mixed DNA profiles is

*Corresponding author. Email: Duncan.Taylor@sa.gov.au

possible and consequently the consideration of peaks just above or just below analytical threshold becomes important.

There have been numerous published methods that describe how the AT could be determined. For a review the reader is referred to the work of Bregu et al.[1]. Some recognise that there are different factors that affect the AT, such as dye colour, input DNA amount or instrument[1,2]. The ideal situation is that these factors are considered on a sample by sample (and even locus by locus) basis and applied to the profile[3]. However, in order to balance the laboratory's sample processing capability with interpretation needs, the laboratories may need to apply a single AT that applies to all profiles, or an AT that is based on dye label, and is set at a level designed to screen out much low level artefactual fluorescence. Thus, it is of value to address the issue of sub-AT information from a standpoint that continues to address the balance between sample processing and interpretation. As such, the purpose of this work is to examine the effects of using a sub-AT threshold signal on interpretation rather than investigating methods to determine the AT. This work considers that no matter where the AT is set, peaks will exist below it that appear allelic and may affect interpretation.

This work evaluates some options for analysts to deal with sub-threshold information and the risks or benefits associated with each in the context of analysis within a continuous DNA interpretation system. We introduce a novel method for dealing with sub-threshold data implemented within the STRmix programme that allows the user to specify a prior belief in mixture proportions.

Much of the discussion will be dominated by the topic of choosing a number of contributors for analysis, which is where the sub-AT peaks will have their biggest impact on interpretations.

There have been various works that look at the consequences of overestimation or underestimation of the number of contributors[4,5]. In general, the consequences of underestimation are that known contributors are excluded due to the forced pairing of peaks that in reality do not pair. The consequence of overestimation is more complex; doing so can have very little effect on a major contributor to a DNA profile and a more marked effect on a minor contributor. This is only true for continuous systems that take peak heights into account. For a semicontinuous system the effect of overestimation will have an effect on all contributors to a mixture as more genotype sets are considered for all contributors to the mixture (see Benschop et al.[6]). There is also a greater number of non-contributors that are given relatively neutral likelihood ratios (*LR*s) as the analysis is accounting for more potential dropout.

The Scientific Working Group on DNA Analysis methods (SWGDAM) guidelines for the validation of probabilistic genotyping systems[7] advise a study of over and underestimation of contributor numbers (at 4.1.6.4) so that the impact of the above-mentioned issues are known for the system being validated. There are methods available that do not require a number of contributors to be assigned[8]; however, the majority of current probabilistic software programmes do require a choice of number of contributors.

This leads to the question of how, if at all, sub-threshold information should be taken into account when making the choice of number of contributors. We consider four broad categories for consideration:

(1) ignore the presence of sub-threshold peaks when interpreting DNA profiles;
(2) change the method by which data are generated (either by lower the AT or carrying out replicate PCRs);

(3) use informed priors on mixture proportion in a probabilistic system;
(4) do not interpret the DNA profile.

1.1. Ignore the presence of sub-threshold peaks when interpreting DNA profiles

To consider the performance and consequence of utilising sub-threshold information when carrying out an interpretation we first start by considering the scope of the issue. We do this in two ways; first via a simulation designed to give an indication of how ignoring sub-threshold information will lead to an underestimate of the number of contributors in the most high-risk situations and, secondly, a demonstration of the practical consequences of ignoring sub-threshold data.

We first start by considering the probability that by ignoring sub-threshold information, a low-level two-person mixture would be assigned as a single source profile. We do this by simulating two contributors with low levels of DNA and different levels of allele sharing and over various analytical thresholds. Twenty-one locus profiles were simulated and the peak heights and AT are intended to be realistic for an Applied Biosystems 3130 capillary electrophoresis (CE) system (Thermo Fisher Scientific, CA). Details of this simulation are given in Appendix 1.

Simulation was chosen in this part of the study because it allows for control over the experimental conditions and for a large number of experiments (for example, Table 1 gives the results of 150,000 simulated mixtures).

Table 1 gives the number of simulations (out of 1000) of two low-level contributors that when combined collectively gave a profile that looked like a single contributor. Simple allele count per locus was used to assign the number of contributors. Use of peak heights is likely to be superior but at such low-levels this is not likely make a significant difference to the count[9].

Inspection of Table 1 suggests that, under the trialled circumstances, there is a high probability of the alleles from two individuals masquerading as a low-level single source profile. The table also shows that this effect is likely to be reduced at lower AT.

This simulation informs the probability of assigning one donor if there are in fact two. It is important not to confuse this with the probability that there are two if we assign one. This latter probability is what we really want. To obtain this probability we need the prior probabilities that there are one or two contributors in a profile. We are allowed to know what type of sample it is and what analysis regime we have employed but we cannot use profile information itself. We will use equal priors for this work, accepting that this was an arbitrary choice. Making this choice will restrict the lower bound probability that a profile is single source, given that it appears as single source to 0.5. Using these priors the probabilities in Table 2 are obtained (details of the calculation appear in Appendix 2).

For the CE system that we are simulating here it is likely that peaks above 30 rfu that have passed expert inspection are all allelic. This suggests that for an AT = 100 or 50 rfu there is a possible strategy of using peaks below the threshold to help improve the assignment of the number of contributors.

These results suggest that ignoring sub-threshold peaks when interpreting low level putatively mixed DNA profiles is likely to lead to underestimation of the number of contributors and thereby has the potential to lead to incorrect interpretations. It is unlikely that a blanket rule to ignore such information would be sustainable. There may be concern that these *in silico* mixtures ignore the effect of stutters. Any stutters misassigned as allelic tends to increase the allele count and hence have no effect at all in the direction of underestimation.

Table 1. The number of simulations (out of 1000) of two low-level contributors that gave a profile that looked like a single contributor based on allele count at 21 loci.

			Average peak height of Contributor 2 (rfu)									
			20	40	60	80	100	120	140	160	180	200
Average peak height of contributor 1 (rfu)	AT = 100 rfu	20	722	734	705	642	549	436	344	230	199	179
		40	734	947	869	718	559	337	194	118	113	78
		60	705	869	746	530	302	119	52	36	17	9
		80	642	718	530	283	95	22	6	3	0	0
		100	549	559	302	95	19	4	0	0	0	0
			Average peak height of Contributor 2 (rfu)									
			10	20	30	40	50	60	70	80	90	100
	AT = 50 rfu	10	754	694	633	557	448	356	249	201	168	137
		20	694	757	520	378	239	122	71	34	33	15
		30	633	520	305	151	57	19	10	2	1	1
		40	557	378	151	70	19	4	0	0	0	1
		50	448	239	57	19	2	0	0	0	0	0
			Average peak height of Contributor 2 (rfu)									
			10	20	30	40	50	60	70	80	90	100
	AT = 30 rfu	10	709	504	315	227	117	71	57	40	30	40
		20	504	302	110	32	16	5	1	0	2	0
		30	315	110	16	1	0	0	0	0	0	0
		40	227	32	1	0	0	0	0	0	0	0
		50	117	16	0	0	0	0	0	0	0	0

Table 2. The probability that the peaks above AT are from a single source (S) given that they look like a single source on simple allele count (AS), Pr(S|AS).

Masking	0.2		0.5	
Mean peak height in range AT (rfu)	10–50 rfu	10–100 rfu	10–50 rfu	10–100 rfu
	Pr(*S*|*AS*)			
30	0.91	0.98	0.70	0.82
50	0.66	0.87	0.56	0.67
100	0.56	0.61	0.59	0.56

We do however look at a number of *in vitro* mixtures. A range of four person mixtures were amplified using GlobalFiler (Thermo Fisher Scientific, CA), as per the manufacturer's instructions. Amplification fragments were resolved using the ABI PRISM 3130xl Genetic Analyser and analysed in GeneMapper ID-X to obtain peak height information for each profile. These mixtures are samples 22 to 31 from Ref. 10, amplified in triplicate except for sample 23 where there were only two replicates, leading to a total of 29 profiles. We reproduce the relevant mixture information from Ref. 10 in Table 3.

Profiles were analysed using ATs of 30 rfu, 50 rfu and 100 rfu. While it is possible to construct simpler mixtures that could be used in this experiment, we chose four-person mixtures due to the high probability that the number of contributors can be underestimated, the higher probability that masking or dropout may occur and as an example of profiles where the use of sub-AT information could have an important impact on the interpretation. Later (in Table 4) we show how, for the data sets used, the number of contributors could be underestimated over half the time.

Profiles were analysed using STRmix V2.3 which utilises models described in Refs 11−13 (exact software settings used are available from the corresponding author on request). In all analyses the Y-indel locus and DYS391 were ignored. A uniform probability for an allelic drop-in of 0.0017 was used (up to 75 rfu) for the 30 rfu and 50 rfu AT and a drop-in probability of zero was used for the 100 rfu AT, in line with laboratory observations.

Two experiments were carried out to investigate the consequences of ignoring the sub-threshold information when determining the number of contributors.

Table 3. Mixture proportions and PCR setup.

Tubes	mixture ratios for contributor C_1:C_2:C_3:C_4	Total DNA added to PCR (pg)
22	1:1:1:1	400
23		200
24		50
25		20
26		10
27	4:3:2:1	400
28		200
29		50
30		20
31		10

Table 4. Assigned number of contributors (based on peak count) are given showing the effect that lowering AT or carrying out replicates has on the ability to determine the number of contributors.

			AT = 10 rfu		AT = 30 rfu		AT = 50 rfu		AT = 100 rfu	
Template (pg)	ratio	replicate	1 PCR	3 PCR	1 PCR	3 PCR	1 PCR	3 PCR	1 PCR	3 PCR
400	1:1:1:1			4		4		4		4
		1	4		4		4		4	
		2	4		4		4		4	
		3	4		4		4		4	
	4:3:2:1	1	4	4	4	4	4	4	4	4
		2	4		4		4		4	
		3	4		4		4		4	
200	1:1:1:1	1	4	4	4	4	4	4	4	4
		2	4		4		4		4	
	4:3:2:1	1	4	4	4	4	4	4	3	3
		2	4		4		4		3	
		3	4		4		4		3	
50	1:1:1:1	1	3	4	3	4	3	3	1	2
		2	4		3		2		1	
		3	4		3		3		2	
	4:3:2:1	1	3	4	3	4	2	3	2	2
		2	4		3		3		2	
		3	3		3		3		2	
20	1:1:1:1	1	3	4	2	2	1	1	0	1
		2	3		2		1		0	
		3	3		2		1		1	
	4:3:2:1	1	2	3	2	2	1	2	0	1
		2	3		1		1		1	
		3	3		2		2		1	
10	1:1:1:1	1	2	3	1	1	1	1	0	0
		2	2		1		1		0	
		3	2		1		1		0	
	4:3:2:1	1	2	3	1	2	1	1	0	0
		2	2		1		0		0	
		3	3		2		1		0	

Experiment 1. Utilising sub-threshold information

First, the correct number of contributors was assigned to each profile during analysis and the *LR*s were calculated using the propositions:

H_p: The person of interest (POI) and three unknown individuals are the sources of DNA.

H_d: Four unknown individuals are the sources of DNA.

The POI was varied to be each of the four known contributors and 186 randomly selected non-contributors. *LR*s were calculated using an in-house self-declared Caucasian GlobalFiler database and using the product rule. This amounts to 116 STRmix analyses compared with known donors and 5394 comparisons to non-donors.

Experiment 2. Ignoring sub-threshold information

In this experiment, the number of contributors was chosen ignoring sub-threshold information i.e. based purely on the number of detected peaks above the varying AT. Using the chosen number of contributors, *N*, *LR*s were calculated using the propositions:

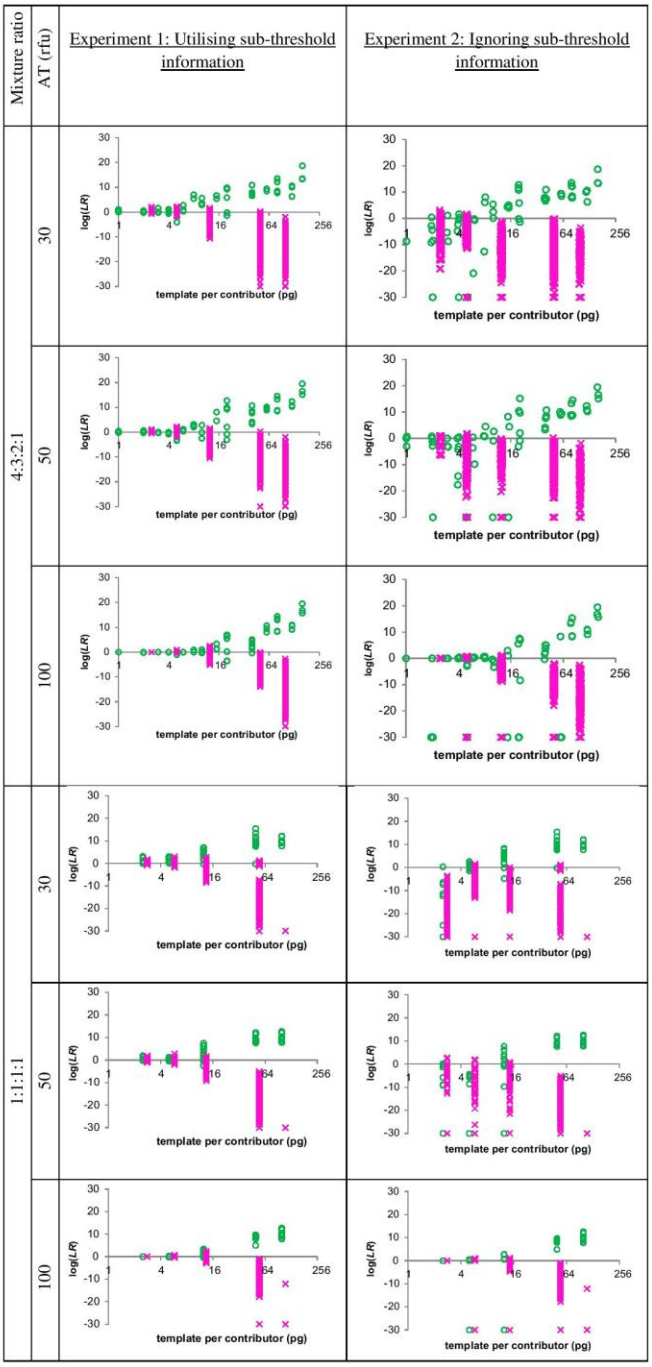


Figure 1. $Log_{10}(LR)$ versus template per contributor (pg) using sub-threshold information (experiment 1) or ignoring sub-threshold information (experiment 2) for a range of four person profiles.

H_p: The POI and (N–1) unknown individuals are the sources of DNA.

H_d: N unknown individuals are the sources of DNA.

The POI was varied to be each of the four known contributors and 186 randomly selected non-contributors. *LR*s were calculated using an in-house self-declared Caucasian GlobalFiler database and using the product rule.

Figure 1 shows the $\log_{10}(LR)$ produced for these comparisons. The *LRs* produced from comparisons to known contributors are signified by a green point and those produced from comparisons to known non-contributors are signified by a pink cross. A minimum value for $\log_{10}(LR)$ of –30 was used, and any *LR*s obtained that fell below this were given the value of –30. The amount of DNA contributed by each known contributor was known from the experimental design. When comparing to non-contributors, the choice of input DNA (for Figure 1) was not known as the non-contributor could align with any of the contributors' input DNA amounts. For known non-contributors the amount of input DNA was assigned as the total amount of DNA added to the PCR divided by the number of contributors. Due to the amount of information present in these graphs we also provide (as supplementary material) the same information but displayed by plotting the $\log_{10}(LR)$ value when considering or ignoring sub-threshold information against each other.

Figure 1 shows that underestimating the number of contributors can cause a $\log_{10}(LR)$ to become less than 0 (sometimes to minimum cap of the graphs) of a true trace contributor in some cases (note the scattered green circles at low log(*LR*) for low template). This is the expected outcome for underestimation[4,5]. We have chosen profiles that are most difficult to interpret due to complexity and high levels of dropout. In addition, a detailed examination of peak heights will be of some but limited use since the donor in dispute is trace and at the limits of the AT. In theory there should be a greater ability to exclude using fewer contributors and this is visible in the results (note the generally lower values for the crosses in the right-hand set of graphs in Figure 1).

This experiment looks at the consequences of underestimation of N and shows that utilising sub-threshold information can partially mitigate the issue. However, use of sub-threshold peaks should be tempered by the relative strength and amount of the putative additional contributor. When assigning a number of contributors based on sub-threshold information there is a risk that an overestimation can occur if any artefacts are considered allelic. It should therefore be balanced by reference to the previously published work[5,14] which showed that an increase in N beyond that required, can alter the *LR* for a true trace contributor and mildly increase the risk of low grade *LR* greater than one.

1.2. Change the method by which data are generated (either by lowering the AT or carrying out replicate PCRs)

To investigate the extent to which generating additional data can assist in interpretation we considered two possible strategies, first a lowering of the AT and second by generating additional PCR replicates. It has already been shown[10] that providing additional, relevant information into the analysis of DNA profile data increases the ability to distinguish a true from a false proposition. We also recognise that due to reasons of practicality there is going to be a limit to which laboratories are willing to lower their AT, and as stated in the introduction, no matter where this level is, there will always be data that appear just below it. We show the effect of lowering the AT as a means to

assist laboratories in their choice of AT, when they will inevitably have to weigh up throughput considerations again data generation.

We analyse the 29 mixed DNA profiles outlined in Table 3 using four different AT (10, 30, 50 and 100 rfu) and considering each of the three PCR replicates individually or in combination in order to determine the number of contributors.

Table 4 shows the effect that lowering AT, using sub-threshold information, or carrying out replicates has on the ability to determine the number of contributors for the data used in this study. For example, inspection of the 1:1:1:1 mixture results at 20 pg individual DNA from Table 4 shows that at AT=50 rfu each of the three individual profiles (1 PCR) appeared to have originated from only one contributor based on allele count. When the AT was reduced to 30 rfu the profiles appeared to have originated from two contributors with more unmasked alleles observed for each contributor. At 10 rfu, when all three replicates are analysed together (3 PCR), the correct assignment of four contributors is made.

1.2.1. Replication

Replication led to some improvement particularly at the fringes when significant portions of the data are dropping out. This can be seen in Table 4 in the 50 pg samples using an AT of 30 rfu, all six of these samples individually detected information that could be described by three individuals, but were clearly four when taking multiple replicates into account. The results in Table 4 also show that amplification can only assist so much. Sticking with an AT of 30 rfu, any samples that were amplified with 10 pg or 20 pg of DNA remained describable by fewer than four individuals even with three replicates. For these samples there is a need to consider what the correct answer is. For example, if the peaks above AT come from three of the four contributors, the 'correct' answer is probably nearer to three rather than four.

There is a resource cost associated with routine repeat amplifications that will need to be considered in forensic laboratories.

1.2.2. Lowering the AT

Comparing graphs vertically in Figure 1 shows very little noticeable improvement in the ability to discriminate true from false donors. However comparing rows horizontally in Table 4 suggests that lowering the AT or using sub-threshold information leads to improved ability to assign the number of contributors. There is a cost in expert time in using very low thresholds. Although no evidence is presented here we assume that at very low thresholds even the most skilled experts will let through artefacts occasionally.

Swaminathan et al.[15] created a continuous method for contributor number assignment (called NOC*It*) and compared this to maximum allele count and maximum likelihood methods. When carrying out the maximum allele count method they found that allowing the AT to shift to the point of baseline noise (19 to 52 rfu) performed worse at estimating the number of contributors than having it fixed at a higher level above baseline noise (50 rfu). While the text does not specifically comment on the reasons for this finding, it may be due to low level artefacts, or stutters appearing above the ratio threshold used being counted as allelic.

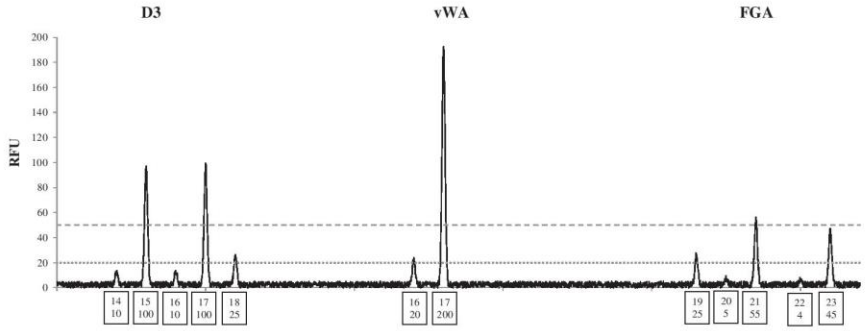


Figure 2. Three loci of a mixed DNA profile with AT shown as a dashed line for 50 rfu and dotted line for 20 rfu. Boxes show peak designation and height.

1.3. Use informed priors on mixture proportion in a probabilistic system

It is possible to provide the analytical system with information that a low level sub-threshold contributor is believed to exist. Consider the mixed DNA profile shown in Figure 2. The known sources of DNA are:

Contributor 1: D3:[15,17], vWA:[17,17], FGA:[21,23]

Contributor 2: D3:[17,18], vWA:[16,18], FGA:[19,19]

In this instance considering the AT as 50 rfu there appears to be a sub-threshold contributor present; however, the detected information present in the profile can be described by a single contributor. Peaks detected at 50 rfu are too weak to be paired with complete certainty at D3 or designated as a homozygote at vWA (using only a single replicate), although their pairing would be the most supported combination. There is therefore likely to be a mild impact of the presence of the sub-threshold peaks on the detected peaks, i.e. the presence of the sub-threshold D3:18 means we would accept a [15,18] or [17,18] pairing for the 'major' some proportion of the time with the 17 or 15 peaks (respectively) coming from a second contributor. The analyst may choose to use the presence of the sub-threshold peaks to consider the profile as originating from two individuals.

We demonstrate the power that providing information, even seemingly minor, can have on the ability of continuous systems to interpret DNA profile data. Before carrying out the experiment there are several predictions that can be made from theory. Consider two *LR*s that could be calculated from these data.

Proposition pair 1

H_{p1}: Contributor 1 and an unknown individual are the sources of DNA.

H_{d1}: Two unknown individuals are the sources of DNA.

Proposition pair 2

H_{p2}: Contributor 2 and an unknown individual are the sources of DNA.

H_{d2}: Two unknown individuals are the sources of DNA.

Table 5. LRs produced for comparison to contributors to epg shown in Figure 2.

		Uniform priors AT = 50 rfu	Using informed priors AT = 50 rfu	Uniform priors AT = 20 rfu
Contributor 1	*LR*	63	108	310
Contributor 2		0.097	0.24	6

If the profile is analysed as a two-person mixture with no guiding information from the analyst even with no significant imbalances in the observed peaks then the analysis will likely split the profile into two roughly equal contributors. Proposition pair 1 will yield an *LR* that favours H_{p1} as most of Contributor 1's peaks are detected, but it will be low as the genotype probability will be spread approximately evenly across a number of genotypes. Proposition pair 2 will yield an *LR* that will likely provide some support for H_{d2} to the profile. The reason for this is that Contributor 2's peaks are not detected and so their presence would have to be explained with multiple dropouts. If the system is supplied with some guiding information that there are two unevenly contributing individuals then we would expect that more weight would be placed on pairing the observed peaks for the major, which we would expect to translate to an *LR* that provides more support for H_{p1} in proposition pair 1. For contributor 2 to be the minor contributor, their peaks have still dropped out; however, now the system is expecting a low template contributor and will be more tolerant of dropout. We therefore would expect the *LR* obtained from proposition pair 2 to be closer to one. Finally, when reading to AT of 20 rfu then more information is given to the system. Informed priors for mixture proportion are no longer required as the information being used to interpret the profile is all being used in the analysis. We would expect a divergence of mixture proportion to be obtained naturally from the data provided and that the *LR* produced from either proposition pair will support the corresponding prosecution proposition.

We now turn to results obtained in practice. The DNA profile in Figure 2 was analysed using STRmix V2.3.06 first using an AT of 50 rfu and providing the system with no information beyond that it has originated from two individuals. Owing to the low peak heights under these circumstances the mixture proportions obtained were 47%:53%.

Secondly the same analysis was carried out in STRmix but supplying mild prior distributions for mixture proportions of $N(0.75, 0.25)$ for contributor 1 and $N(0.25, 0.25)$ for contributor 2. We use priors on the mixture proportion; however, we realise that it is in fact the template DNA amount that these priors will be acting on. Priors for mixture proportions are displayed for the ease of the user because doing so does not need them to consider how other effects within the DNA profile such as degradation and locus specific amplification efficiencies interact with the template to generate peak heights. Mixture proportions will automatically scale with peak intensity and so the user does not need to scale their priors for each similarly proportioned mixture. We also recognise that Gaussian distributions extend beyond the interval [0,1] but only apply them within this range.

The mean of the posterior for mixture proportions from the analysis were 85%:15%. The third analysis was for data using AT of 20 rfu, and not providing informed priors for mixture proportion. This time the mean of the posterior for mixture proportions from the analysis were 79%:21%. The *LR*s when comparing contributors to the three analyses can be seen in Table 5. The trend of *LR*s fits what is expected by

theory and demonstrates the point that even just supplying the information that the analyst has a prior belief in the mixture proportions based on sub-threshold data (without supplying that specific data to the analysis system) aids in the analysis and produces a result that is more intuitively aligned with the human assessment.

1.4. Do not interpret the DNA profile

At the laboratory at Forensic Science South Australia, an audit of samples received over a one-month period revealed that 54% of samples fell into what is classically called transfer or contact DNA and 34% of samples yielded a total DNA concentration of less than 10 pg/μL. There would be many more that would possess less than this level for individual contributors to mixed samples. These profiles are likely to suffer from significant allelic dropout and be within the range where sub-threshold information will be present.

A simple solution to the problems of interpreting epgs with sub-threshold peaks might be to deem all such profiles as too complex; however, given the portion of profiles that this group would represent it is unlikely this would be a sustainable practice. We do not mean this to be an excuse to interpret poor quality data, quite the contrary, instead we mean this statement to highlight the need to determine what data can be interpreted (which we hope we have started in this work).

The question must be asked whether certain profiles *should* be analysed. This is a different question to whether a profile *can* be analysed. Taking a position of theoretical purity, all data can be analysed as long as models exist to describe it. As the information content of the data decreases, or the uncertainty surrounding the interpreted profile increases, there will be an inevitable drop in the discriminating power the model will provide using the data. This is the desired behaviour and correctly represents the strength of the data. There is no limit to which this thinking can be applied. For example, the models already exist that an analyst could obtain an epg that exhibits a single weak peak of putative artefactual status and choose to analyse it, considering it may originate from anywhere between one and five individuals. After what is likely to be several hours of processing and analysis, utilising highly complex statistical, mathematical and biological theory and being provided with many pages of detailed output the interpretation system would no doubt inform the analyst of what they already knew, there is no information in the datum to discriminate true from false propositions.

Whether something should be analysed will depend on a number of factors, many of which will not directly relate to the epg in question. Ultimately it will be a decision made by the analyst that the potential discriminating power that epg could provide, in the context of the case and laboratory environment, is worth the interpretation and analysis time.

2. Interpretation of putative stutter peaks

When interpreting a DNA profile that has a major component and one or more minor components that are in the same peak height range as stutter of the major, then some assessment of the nature of small peaks in stutter positions will need to be made by the analyst.

It is worth discussing the 2006 ISFG[16] Recommendation 6, which states:

> If the crime profile is a major/minor mixture, where minor alleles are the same size (height or area) as stutters of major alleles, then stutters and minor alleles are indistinguishable.

> Under these circumstances alleles in stutter positions that do not support H_p should be included in the assessment.

It is the authors' experience that this statement is sometimes taken as meaning 'all peaks in stutter positions must be treated as allelic' as it has been used as such for interpretational attack in court. We suggest that this is not the intent of the authors of Ref. 16 when making this recommendation. In the same publication, the preceding sentence gives an example of when the recommendation would have an effect, and states that under those circumstances '…the probability of stutter must be considered…'. Probabilistic systems take into account the ambiguous nature of peaks by calculating the probability of that peak if it is purely stutter as opposed to it being partially allelic (given a number of parameters dealing with intrinsic properties of the DNA profile such as DNA amounts, degradation, genotype sets, etc.). Sometimes the choice of number of contributors will mean that the certain peaks within the profile will be considered unambiguously as entirely stutter, however this is a perfectly acceptable outcome. To consider all peaks in stutter positions as allelic would see an overestimation of the number of contributors in a large proportion of samples and would be against the ethos that each party is allowed its best explanation of the evidence.

This leaves the analyst with the task of making an assessment of the nature of peaks in stutter positions as to their status. There is a risk here of either overestimating or underestimating the number of contributors to the profile and we point the reader to Refs 4 and 5 for the outcomes of either of these eventualities when using a continuous system including examples of ambiguous stutter peaks. Our intention in this paper is not to trial or recommend methods for dealing with ambiguous peaks in stutter positions and we do not do so. All we suggest is that the method used should take into account known stutter values for alleles/loci and the profile should be considered holistically, which may include an assessment of the presence of peaks below the AT.

3. Conclusion

Continuous systems (at least STRmix as trialled here) can overcome the issues of missing low-level data with minimal effects on the outcome of the analysis. The effects of overestimation of the number of contributors may not be too severe as long as the system has been reliably validated for this policy. This situation should not be used to enable a reduction of valid quality practices such as replication and careful expert inspection of profiles and cannot be assumed to be conservative. However, any system, even one possessing the soundest theoretical basis, that cannot withstand the rigours of practical use, is destined to remain nothing more than a nice idea. We have discussed strategies to mitigate the effect of uncertainty in the number of trace contributors present when sub-threshold information is present in a DNA profile. We support replication and lowering the AT whenever practical. The use of sub-threshold data without lowering the AT may be useful in some cases. The effects of mis-assignment of N in either direction are relatively mild and restricted to *LR*s less than one when comparing known contributors and low *LR*s greater than one when comparing known non-contributors.

We believe that treating the number of contributors as an unknown nuisance variable is the best long-term solution. An even better solution would be to combine the treatment of number of contributors as a nuisance variable with an expert system that utilises fluorescent signal directly and has models for different known artefacts. In such a system all data would be treated probabilistically and the tyranny of thresholds would

be completely abolished. We are not aware of any system that can perform at this level and so can provide no examples of how it would perform.

Last, we suggest that some profiles are simply too complex and should not be interpreted. Ultimately it is the role of the scientist to assess each profile on its own merits and the case context in order to determine if and how analysis will proceed.

Acknowledgements

We thank Lisa Melia and Johanna Veth and two anonymous reviewers for their valuable comments that greatly improved this manuscript.

Funding

This work was supported in part by grant 2014-DN-BX-K028 from the US National Institute of Justice. Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the US Department of Justice or Commerce. Certain commercial equipment, instruments, or materials (or suppliers, or software) are identified in this paper to foster understanding. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

References

1. Bregu J, Conklin D, Coronado E, Terrill M, Cotton R, Grgicak C. Analytical thresholds and sensitivity: establishing RFU thresholds for forensic DNA analysis. J Forensic Sci. 2012;58 (1):120–129.
2. Taylor D, Bright J-A, McGoven C, Hefford C, Kalafut T, Buckleton J. Validating multiplexes for use in conjunction with modern interpretation strategies. Forensic Sci Int Genet. 2015;20:6–19 doi:10.1016/j.fsigen.2015.09.011.
3. Gilder JR, Doom TE, Inman K, Crim M, Krane DE. Run-specific limits of detection and quantitation for STR-based DNA typing. J Forensic Sci. 2006;52(1):97–101.
4. Bright J-A, Curran J, Buckleton J. The effect of the uncertainty in the number of contributors to mixed DNA profiles on profile interpretation. Forensic Sci Int Genet. 2014;12:208–214.
5. Bright J-A, Taylor D, Curran J, Buckleton J. Searching mixed DNA profiles directly against profile databases. Forensic Sci Int Genet. 2014;9:102–110.
6. Benschop C, Haned H, Jeurissen L, Gill P, Sijen T. The effect of varying the number of contributors on likelihood ratios for complex DNA mixtures. Forensic Sci Int Genet. 2015;19:92–99.
7. Scientific Working Group on DNA Analysis Methods (SWGDAM). Guidelines for the validation of probabilistic genotyping systems, 2015.
8. Taylor D, Bright J-A, Buckleton J. Interpreting forensic DNA profiling evidence without specifying the number of contributors. Forensic Sci Int Genet. 2014;13:269–280.
9. Taylor D, Buckleton J. Do low template DNA profiles have useful quantitative data? Forensic Sci Int Genet. 2015;16:13–16.
10. Taylor D. Using continuous DNA interpretation methods to revisit likelihood ratio behaviour. Forensic Sci Int Genet. 2014;11:144–153.
11. Taylor D, Bright J-A, Buckleton J. The interpretation of single source and mixed DNA profiles. Forensic Sci Int Genet. 2013;7(5):516–528.
12. Bright J-A, Taylor D, Curran JM, Buckleton JS. Developing allelic and stutter peak height models for a continuous method of DNA interpretation. Forensic Sci Int Genet. 2013;7 (2):296–304.
13. Bright J-A, Taylor D, J.M. C, Buckleton JS. Degradation of forensic DNA profiles. Aust J Forensic Sci. 2013;45(4):445–449.
14. Bright J-A, Curran JM, Buckleton JS. The effect of the uncertainty in the number of contributors to mixed DNA profiles on profile interpretation. Forensic Sci Int Genet. 2014;12:208–214.

15. Swaminathan H, Grgicak CM, Medard M, Lun DS. NOCIt: a computational method to infer the number of contributors to DNA samples analyzed by STR genotyping. Forensic Sci Int Genet. 2015;16:172–180.
16. Gill P, Brenner CH, Buckleton JS, Carracedo A, Krawczak M, Mayr WR, Morling N, Prinz M, Schneider PM, Weir BS. DNA commission of the international society of forensic genetics: recommendations on the interpretation of mixtures. Forensic Sci Int. 2006;160:90–101.

Appendix 1

Peaks for each of the two contributors were simulated from a lognormal distribution with mean μ and variance $\frac{4}{\mu}$. With probability 0.2 a peak was masked. Masking can be thought of as happening because a major contributor is present or because the two traces mask each other. The number of peaks per locus was counted and any profile that had only 0–2 peaks per locus was checked to see that it did have contributions from each contributor. This is the number of profiles out of the 1000 simulations appearing in Table 1.

Appendix 2

Let

S be the event that the peaks above AT come from a single source;

T be the event that the peaks above AT come from two sources;

AS be the event that the peaks above AT appear to come from a single source by simple allele count.

Values for the mean μ were drawn from either U[10,50 rfu] and U[10,100 rfu] for each of the two contributors. $\Pr(AS|S)$ and $\Pr(AS|T)$ were calculated using the simulation described in Appendix 1 (1000 simulations were used). Masking was set at 0.2 and 0.5.

The desired probability was obtained as:

$$\Pr(S|AS) = \frac{\Pr(AS|S)\Pr(S)}{\Pr(AS|S)\Pr(S) + \Pr(AS|T)\Pr(T)}$$

and assuming $\Pr(S) = \Pr(T)$. These values appear in Table 2.

Chapter 7: Extending the theory in the future

Chapter 7 considers how the theories of DNA profile deconvolution and evaluation presented in this thesis can be extended into the future. There are two broad groups of such considerations. First, are enhancements to the current models, either through refining the existing models (as given in chapter 2, for example the refinement of the stutter model described in 2.6, to the LUS stutter model described in 2.2, to the multi-LUS model described in 2.5), or modelling new factors that affect DNA profile generation or behaviour. There is a trade-off with such enhancements between the complexity of the model and the amount of peak height information that is explained. If the model is simple then the subtleties of small deviations from expected fluorescence will be unnoticed and their ability to distinguish between explanations lost, because the effects of multiple real-world events are being described by a single model. However, a simple enough model can be recreated by hand, understood by all, and run in seconds.

If the model is complex then the subtleties of the multiple interacting real-world events will be explained, and the ability to distinguish between competing explanations at its peak, however there are associated costs. One is the cost of comprehensibility. A highly complex system will be understood by less people (or just less understood by people), which has the disadvantages of the acceptance first by the scientific community and second by the legal community. Also, as a system becomes more complex it, by necessity, will take more computing power and longer to run. Another potential issue is that as systems become more complex and the number of interacting models increases, there is a tendency for systems to become too forgiving, i.e. the values of parameters within the models can shift to positions that can describe nonsensical data, rather than simply indicating that there is something wrong with the input. As an example, there are certain chemicals that, when present in a PCR, will inhibit and retard the amplification of DNA fragments, to different degrees for different targeted regions. The result is a DNA profile that does not have the expected 'ski-slope' pattern with respect to molecular weight. STRmix™ could be extended by adding 'inhibition' as a model within the system, that would allow the peaks within the profile(s) to shift the tolerance of the system to extreme amplification efficiencies. But the wider question is whether it is better to fix the biological issue of PCR inhibition first, rather than attempting to deal with it statistically. A 'good data in' ethos is one to which the forensic community fully subscribe and so models have not been added that are designed to deal with clearly substandard data. A balance must be struck between the benefits and drawback of refining models too far.

The second broad group of considerations is in the extension of the current theory to apply to new situations. There are two areas where publications are provided in this thesis:

1) The extension of the theory to apply to Y-STR data
2) The extension of the theory to account for uncertainty in the number of contributors

While the theory of these two extensions has been explored and published, neither is yet in active casework use. The reasons for this lack of application are quite different for the two situations and are explained in the following sections.

7.1 YSTR extension

In certain scenarios it is advantageous to target male DNA specifically (most commonly in rape scenarios). The type of DNA (that is STRs) tested in Y-chromosome profiling is the same as in autosomal profiling kits. Y-STR profiles also have very similar behaviours in that template, degradation, stutters and amplification efficiencies should all occur in the same manner. This was shown to be true in the publication in this section, where a deconvolution model was adapted to deal with Y-STR data and performed to much the same high quality as on autosomal data. The motivation behind the paper in this section was not directly to build a deconvolution tool for Y-STR data, but rather an attempt to use the knowledge of DNA profile behaviour obtained from developing STRmix™ to develop probabilistic interpretation guidelines for Y-STR profiles. In other words, because no continuous system of DNA profile interpretation exists for Y-STRs, laboratories are forced to develop threshold-based guidelines. These suffer from all the drawbacks mentioned in the introduction. The paper in this section was an attempt to address these issues (partially at least) using continuous theory.

The broader reason as to why no continuous interpretation exists for Y-STRS is what comes after the deconvolution. Because the Y-chromosome is inherited from father to son in an unaltered block (apart from when mutation occurs) the classic form of the LR that deals with Y-STR data uses whole-profile haplotype frequencies. The problem with this is that modern kits possess 20 or more loci and consequently an astronomical number of whole-profile haplotype frequencies are required when dealing with mixed samples. So intractable is the impasse that unless a locus-by-locus approach can be developed, which performs well under the many imperfections present in human populations, complex mixtures of Y-STRs simply cannot be evaluated. Work is ongoing to address this issue.

Manuscript: Using probabilistic theory to develop interpretation guidelines for Y-STR profiles. D Taylor, JA Bright, J Buckleton. (2016) Forensic Science International: Genetics 21, 22-34 – *uncited*

Statement of novelty: This work takes existing theory from continuous interpretation of Autosomal STR data and applies it to Y-STR profile data to develop guidelines for interpretation (in the absence of an existing fully continuous method of analysing Y-STR profiles).

My contribution: Main author and sole simulation programmer. Main contributor to theory.

Research Design / Data Collection / Writing and Editing = 70% / 100% / 60%

Additional comments:

Forensic Science International: Genetics 21 (2016) 22–34

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

Research paper

Using probabilistic theory to develop interpretation guidelines for Y-STR profiles

Duncan Taylor[a,b,*], Jo-Anne Bright[c], John Buckleton[c,d]

[a] Forensic Science South Australia, 21 Divett Place, Adelaide, SA 5000, Australia
[b] School of Biological Sciences, Flinders University, GPO Box 2100, Adelaide, SA 5001, Australia
[c] ESR, Private Bag 92021, Auckland 1142, New Zealand
[d] National Institute of Standards and Technology, 100 Bureau Drive, MS 8980 and 8314, Gaithersburg, MD 20899, United States

ARTICLE INFO

Article history:
Received 27 September 2015
Received in revised form 6 November 2015
Accepted 25 November 2015
Available online 28 November 2015

Keywords:
Y-STR
Profile
Interpretation
Thresholds
Probabilistic interpretation
Guidelines

ABSTRACT

Y-STR profiling makes up a small but important proportion of forensic DNA casework. Often Y-STR profiles are used when autosomal profiling has failed to yield an informative result. Consequently Y-STR profiles are often from the most challenging samples. In addition to these points, Y-STR loci are linked, meaning that evaluation of haplotype probabilities are either based on overly simplified counting methods or computationally costly genetic models, neither of which extend well to the evaluation of mixed Y-STR data. For all of these reasons Y-STR data analysis has not seen the same advances as autosomal STR data. We present here a probabilistic model for the interpretation of Y-STR data. Due to the fact that probabilistic systems for Y-STR data are still some way from reaching active casework, we also describe how data can be analysed in a continuous way to generate interpretational thresholds and guidelines.

1. Introduction

The last few years has seen substantial advances in the ability to interpret, analyse and evaluate autosomal STR data [1–7]. Limited advances have been made for the interpretation of Y chromosome STR profiles, although some investigation into stutter has appeared [8,9]. Y chromosome work makes up a low proportion of casework and presents difficulties with the evaluation of the data. The difficulties with the evaluation of Y-STR data can be broken into two parts; firstly the lack of probabilistic modelling that has been applied to Y-STR data and secondly the difficulties associated with the evaluation of the comparison of Y-STR reference profiles/s to an evidence profile. We suspect that the difficulties in the latter are largely responsible for the lack of work on the former. In this work we deal with the probabilistic modelling of Y-STR data, for which much can be borrowed from already existing autosomal models.

This work is broken into two sections. The first of these is the deconvolution of a profile into its contributing haplotypes using probabilistic theory. We borrow from autosomal models that take into account factors that affect DNA profile behaviour such as template DNA amount, degradation, and locus specific amplification efficiencies. These models can be used to determine the probability of the observed Y-STR profile given potential contributing haplotypes. In doing this, the models allow a 'weight' to be given to each potential contributor haplotype set that acts as an indication of how well the proposed haplotypes describe the observed data. This then lends itself to development of interpretational guidelines.

As there is currently no tool available that can be used to analyse Y-STR data in the manner described within this paper the second section describes how the continuous thinking can be applied to assist in the creation of mixture interpretation guidelines and thresholds such as a stochastic threshold (ST).

1.1. Deconvolution of Y-STR data

1.1.1. Building an expected profile

The deconvolution of autosomal STR data relies on a number of models that are used to describe various properties of DNA profile behaviour. Ultimately, given some parameters we wish to describe the fluorescence observed in one or more electropherograms (epg). In this work we follow the autosomal model of Ref. [1] in that we consider the following 'Mass' parameters to describe observed fluorescence:

* Corresponding author at: Forensic Science South Australia, 21 Divett Place, Adelaide, SA 5000, Australia. Fax: +61 8 8226 7777.
E-mail address: Duncan.Taylor@sa.gov.au (D. Taylor).

http://dx.doi.org/10.1016/j.fsigen.2015.11.010
1872-4973/

1. Template amount, t_n, for each of the n contributors. Collectively we refer to the vector of all template values as T.
2. Degradation, d_n, which models the decay with respect to molecular weight (m) in template for each of the contributors. We constrain degradation to be negative. Collectively we refer to the vector of all degradation values as D.
3. Amplification efficiency, A^l to allow for the observed amplification levels of each locus l. Amplification efficiencies are modelled by a normal distribution, $\log_{10}(A^l) \sim N(0, \sigma^2)$. Collectively we refer to the vector of all amplification efficiency values as A.
4. A replicate amplification multiplier R_r. This effectively scales all peaks up or down between replicates r . We constrain $\sum_r \log(R_r) = 0$. Collectively we refer to the vector of all replication amplification multiplier values as R.

Together we consider D,A, R and T as the mass parameters, M.

The total allelic product, T^l_{anr}, for an allele, a, at locus l, from contributor n in replicate r is modelled as

$$T^l_{anr} = A^l R_r t_n e^{d_n \times m^l_a} X^l_{an} \quad (1)$$

m^l_a is the molecular weight of allele a at locus l. Notice that this is the same formulation for total allelic product for autosomal STR data as described in Ref. [1], except that 'dose' (a term designated as X^l_{an} in Ref. [1], which took into account whether the individual was heterozygous or homozygous for allele a) can only takes values of 0 or 1 depending on whether contributor n contains allele a at locus l. The model we present, at this stage, only considers data for which the primer pair amplifies a single amplicon.

The total allelic product from an allele is split between back stutter, forward stutter and allelic peak. We model back stutter ratio, $\overleftarrow{\pi}^l_a$, and forward stutter ratio, $\vec{\pi}^l_a$, for allele a at locus l using models described in [10,11]. The height of the allelic and stutter peaks formed from allele a are therefore:

$$E^l_{anr} = \frac{T^l_{anr}}{1 + \overleftarrow{\pi}^l_a + \vec{\pi}^l_a} \quad (2a)$$

$$E^l_{(a-1)r} = \overleftarrow{\pi}^l_a O^l_{ar} \quad (2b)$$

$$E^l_{(a+1)r} = \vec{\pi}^l_a O^l_{ar} \quad (2c)$$

where $a-1$ signifies the back stutter product and $a+1$ signifies the forward stutter product. Note that we base the expected height of stutter products on the observed parent peak height rather than the expected parent peak height and so they do not possess a contributor term.

Given the mass parameters the height of allelic peaks are expected to be independent within and between loci. Stutter peaks are dependent on their parent peak heights, but given this they are also expected to be independent within and between loci. This allows the deconvolution to occur in a locus by locus manner as described in Ref. [1], rather than having to consider the entire haplotype or haplotypic mixture as a whole entity. In this way then the deconvolution of Y-STR data becomes very similar to that of autosomal STR data, with the exception that, in our simplified model, a single allele is always expected at each locus from each contributor, rather than there being a possibility of one or two alleles being donated.

Given a set of values for mass parameters, M, and a haplotype for each of the n contributors then an expected profile (which we call E, and is made up of all individual E^l_{anr} terms) can be built by summing the expected allelic peaks from the individual contributors and any back or forward stutters from alleles that differ by a single repeat unit.

$$E^l_{ar} = \overleftarrow{\pi}^l_{(a+1)} O^l_{(a+1)r} + \vec{\pi}^l_{(a-1)} O^l_{(a-1)r} + \sum_n E^l_{anr}$$

$$E \equiv E^1_{1,1} \ldots E^l_{A,R}$$

1.1.2. Peak height variability

There exists a level of peak height variability within STR DNA profiles. We note that the expected peak heights (E) generated from the mass parameters may be different from the observed peak heights (O) in the epgs(s). In large part this is due to variation in the sampling of DNA molecules in a DNA extract for inclusion in PCR reaction [12]. As a base for modelling peak height variability we follow Refs. [1,11]. We extend that theory here so that the difference in an observed peak from its expected height $Pr(O^l_{ar}|E^l_{ar})$ is modelled by Eqs. (3a), (3b) and (3c). In these equations the terms O^l_{ar} and E^l_{ar} on the left hand side of the equations refer to the peak being examined and the '+1' or '−1' on the right hand side of the equations signify peaks relative to the peak in question. Also note that Eq. (3a) and (3b) are the same but we list both to assist in the understanding of Eq. (3d).

$$\log\left(\frac{O^l_{ar}}{E^l_{ar}}\right) \sim N\left(0, \frac{c^2}{E^l_{ar}}\right) \text{for purely allelic peaks} \quad (3a)$$

$$\log\left(\frac{O^l_{ar}}{E^l_{ar}}\right) \sim N\left(0, \frac{c^2}{E^l_{ar}}\right) \text{for purely forward stutter peaks,} \quad (3b)$$

$$\log\left(\frac{O^l_{ar}}{E^l_{ar}}\right) \sim N\left(0, \frac{k^2}{O^l_{(a+1)r}}\right) \text{for purely back stutter peaks and} \quad (3c)$$

$$\log\left(\frac{O^l_{ar}}{E^l_{ar}}\right) \sim N\left(0, \frac{\mathbb{S}^l_{ar} k^2}{O^l_{(a+1)r}} + \frac{\left(\mathbb{A}^l_{ar} + \mathbb{F}^l_{ar}\right) c^2}{E^l_{ar}}\right) \text{for composite peaks} \quad (3d)$$

where $\mathbb{A}^l_{ar}$, $\mathbb{S}^l_{ar}$and $\mathbb{F}^l_{ar}$ are allelic, back stutter proportion and forward stutter proportions of the peak, so that $\mathbb{A}^l_{ar}+\mathbb{S}^l_{ar}+\mathbb{F}^l_{ar} = 1$.

The values of k^2 and c^2 in autosomal work have prior gamma distributions and the point values for these terms is a parameter within the model, so that the analysis can adjust to account for profiles that are more or less variable than the 'average variability' (see Ref. [13] for work describing the nature of the k^2 and c^2 terms and factors that affect their prior distribution).

For autosomal STR analysis the values of k^2 and c^2 have separate prior distributions, $\Gamma(\alpha_1,\beta_1)$ and $\Gamma(\alpha_2,\beta_2)$, respectively, and are independent parameters within the model. The work in Ref. [13] describes how Markov Chain Monte Carlo (MCMC) can be used to provide the peak height variability framework described above, and profiles of known source to obtain the peak height variance constant prior distributions. This works in part because there are two allelic peaks expected at each heterozygous locus, and these peaks act as intra-locus calibrators of each other's expected heights. In the case of haplotypic markers, such a calibration does not exist. The consequence within the MCMC of not having an

intra-locus calibrator could be that a disproportionate amount of peak height variability is assigned to either allelic or stutter peaks (i.e. the model could explain all variability exists in allelic peak heights and that stutter peaks have no peak height variability). To counter this we restrict the evaluation of prior distributions of variance constants to be equal, $\Gamma(\alpha_1,\beta_1)=\Gamma(\alpha_2,\beta_2)=\Gamma(\alpha,\beta)$. In autosomal work this assumption is found to be approximately true depending on the resolution of the stutter model in use (data not shown).

As long as $\Gamma(\alpha,\beta)$ can be determined this now provides a means to assess the differences between expected and observed peak heights within Y-STR profiles.

1.1.3. Haplotypic weight

As mentioned in Section 1.1.1, with given mass parameters the expected height of peaks within and between loci are independent. We follow the work of Ref. [1] by describing the weight (w_j^l) for a proposed genotype set j at locus l (S_j^l) as the probability of the observed data given the genotype set and integrated across the mass parameters:

$$w_j^l = \int_M \prod_a \prod_r Pr(O_{ar}^l | S_j^l, M) Pr(M) \mathrm{d}M \tag{4}$$

And that the weight for the entire profile haplotype is obtained by multiplication of the weights at each locus:

$$w_j = \prod_l w_j^l \tag{5}$$

which is an approximation for the integral that considers whole profile weights as integrals across the whole profile, rather than locus by locus. This approximation has been shown to be valid [1].

We use MCMC in the method described by Ref. [1] to evaluate Eq. (4). In brief, we change all mass parameters and the genotype at one randomly chosen locus within the MCMC at each iteration. The acceptance or rejection of the proposed values is decided by the Metropolis–Hastings algorithm and the weights for each genotype set become the proportion of iterations that contain them.

1.2. Determining the peak height variability of Y-STR data

We use Yfiler® Plus (Life Technologies, CA) as an example for determining peak height variability. The following settings were used:

- Saturation: 7000 rfu determined as per Section 3.7 of Ref. [10].
- Analytical threshold: 30 rfu determined as per appendix 1 of Ref. [10].
- Back stutter ratio: regressions of allele vs back stutter ratios were used in accordance with stutter ratio results obtained from Life Technologies Yfiler® Plus manual [14] chapter 5, Figs. 11–17. The table of regression values used are given in Table 1.
- Forward stutter ratio: a profile-wide forward stutter ratio of 0.01 was used.

DNA from two male individuals was extracted twice using Chelex (Bio-Rad Laboratories, CA) and twice using DNA IQ (Promega Corporation, WI) and amplified at 500, 400, 100, 50, 20, 10, 5 and 1 pg in triplicate, giving a total of 96 samples. All samples containing saturated data or fewer than 10 data points were excluded from the dataset, giving total of 67 samples. This group of 67 samples were analysed using the method described in Section 1 of Ref. [13]. The distribution for the peak height variance constant (not shown) has mean of approximately 9 and a mode of approximately 6. There is some arbitrariness as to choosing a point value from a distribution with which to develop thresholds, and an argument could be made to choose either the mode or the mean, or some other quantile. We have chosen the mean. This value can be 'sanity checked' by comparison with stochastic effects seen in observed data. This is carried out in a manner similar to that described in Section 3.4 in Ref. [10], however instead of considering heterozygous imbalance we graph $\log(O_{ar}^l/E_{ar}^l)$ against E_{ar}^l for allelic peaks. The expected height for E_{ar}^l is obtained by the expected back stutter ratio and observed back stutter peak height:

$$E_{ar}^l = \frac{O_{(a-1)r}^l}{\overleftarrow{\pi}_a^l} \tag{6}$$

Doing so for the 96 samples in the Yfiler Plus dataset produces the graph seen in Fig. 1, where the dotted lines represent the 95% bounds on expected variability (see Ref. [10] for an explanation of the origin of these bounds) and the dashed line represents the approximate bound on the observed data based on an analytical threshold of 30 rfu (approximate because it uses an average value of $\overleftarrow{\pi}_a^l = 0.09$, which is the average value for all peaks in the dataset).

Table 1
Back stutter ratio regression results.

Locus	In regression $\overleftarrow{\pi}_a^l = \beta_0 + \beta_1 a$ value below are: β_0,β_1
DYS576	−0.0311, 0.0078
DYS389I	−0.0700, 0.0100
DYS635	−0.0922, 0.0078
DYS389II	−0.1600, 0.0100
DYS627	−0.0700, 0.0080
DYS460	−0.0525, 0.0125
DYS458	−0.0467, 0.0089
DYS19	−0.1300, 0.0150
YGATAH4	−0.0860, 0.0140
DYS448	−0.0275, 0.0025
DYS391	−0.0400, 0.0100
DYS456	−0.0900, 0.0133
DYS390	−0.2167, 0.0133
DYS438	−0.0200, 0.0057
DYS392	−0.0550, 0.0125
DYS518	−0.1220, 0.0080
DYS570	−0.0133, 0.0067
DYS437	−0.1000, 0.0100
DYS385	−0.0400, 0.0100
DYS449	−0.1300, 0.0100
DYS393	−0.0583, 0.0117
DYS439	−0.0500, 0.0100
DYS481	−0.0844, 0.0122
DYF387S1	−0.2280, 0.0090
DYS533	−0.0400, 0.0100

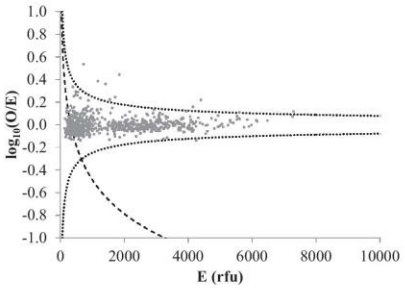


Fig. 1. Observed peak height variability showing 95% bounds (dotted lines) and bound on observed data (dashed line).

Because of the bound on observations it is difficult to assess the coverage of the expected 95% bounds on $\log(O^l_{ar}/E^l_{ar})$. It would be expected that 2.5% of data falls above the upper bound dotted line, and should not be affected by the bound on observations. 1.5% of observations seen in Fig. 1 fall above the upper bound 95% interval, which represents a reasonable alignment with theoretical expectations.

As described in Ref. [13] the distribution of locus amplification efficiency variances over a number of samples can be modelled during a component-wise MCMC using an exponential distribution (termed in Ref. [13] as a hyper-distribution). Doing so for the dataset described here produced an exponential distribution with mean value 0.03. We then use this mean value as σ^2 in the locus amplification efficiency model by $\log_{10}(A^l) \sim N(0,0.03)$.

2. Developing Y-STR guidelines from probabilistic assessment

Without access to probabilistic software that can apply a continuous interpretation method to Y-STR profiles, laboratories will require some rules or guidelines to assist with a 'manual' interpretation. Typically, such guidelines will include:

- An analytical threshold.
- A saturation threshold.
- A stochastic threshold (sometimes referred to as a dropout or homozygote threshold).
- Mixture ratio/proportion guidelines.

The analytical threshold and saturation threshold for Y-STR profiling systems can be determined in the same manner as for autosomal STR systems.

2.1. Stochastic threshold (ST)

It is worth considering whether the term 'stochastic/dropout/homozygote threshold' has meaning in the context of Y-STR profiles. Classically for autosomal STR profiles this denotes a height at which, if we see a single allelic peak, (and given the level of peak height variability for that laboratory/process) we would have a defined level of confidence that it does not possess a partner that had fallen below the analytical threshold. In the context of Y-STR data the same meaning cannot be ascribed as peaks within a locus do not have partners (even loci that have apparent multiple alleles are really amplifications of several different loci from the one primer pair). For single source Y-STR data a ST is more useful to distinguish an allele that has dropped out from a null allele. This by necessity is based on the heights of allelic peaks at other loci. In mixtures the consideration of dropouts (or distinguishing dropout from null alleles) has consequences for haplotype interpretation. Andersen et al. [15] consider dropout probabilities and locus amplification efficiency in a model that includes the molecular weight of the locus in question in order to consider whether dropout or null allele is more likely. We use this same concept, but apply our models for peak height variability and locus amplification efficiency. We use the same probability of a null allele, $\Pr(N) = 0.0002$ as used in Ref. [15].

Therefore we consider two events that could cause an apparent dropout:

1) Underrepresentation of the allele during sampling the DNA extract for PCR.
2) Poor amplification efficiency of the locus.

We seek two values; firstly, the probability of dropout, of a peak with height E^l_{ar} (which we approximate as the average peak height of allelic peaks in the profile for single source profiles), and secondly, the point at which the probability that a peak has dropped out becomes more likely than the probability that the absence of an allele at that locus is due to the presence of a null allele. Let the haplotype with dropout be S_1 and the haplotype with a silent allele be S_2. Formally the probability of the data given the haplotype is $\int_M \prod_l \prod_a \prod_r Pr(O^l_{ar}|S^l_j, M)Pr(M)\mathrm{d}M$ which we can approximate with Eqs. (4) and (5). Let us imagine that the locus with potential dropout is x then the probability of the data for each haplotype is:

$$\int_M \prod_l \prod_a \prod_r Pr(O^l_{ar}|S^l_1, M)Pr(M)\mathrm{d}M \text{ for haplot type 1 and} \quad (7a)$$

$$Pr(N) \times \int_M \prod_{l\neq x} \prod_a \prod_r Pr(O^l_{ar}|S^l_1, M)Pr(M)\mathrm{d}M \text{ for haplot type 2} \quad (7b)$$

We simplify Eqs. (7a) and (7b) by considering that the mass parameters template and degradation provide an expected height, E^*, before amplification efficiency has been considered. By doing this we simplify the problem to consider just the locus where the potential dropout/null has occurred. Therefore to calculate the

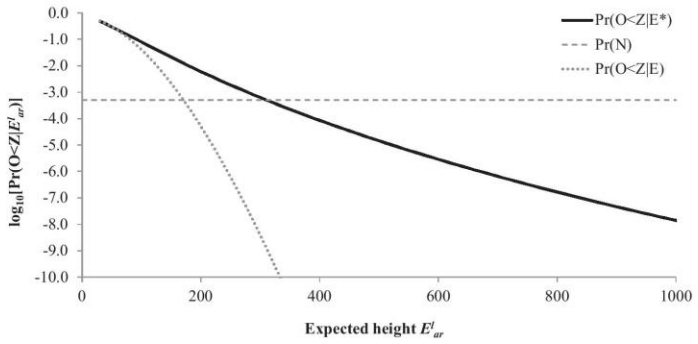


Fig. 2. Probability of dropout (solid line) for expected heights E^l_{ar} when $Z = 30$ rfu. The dashed line is Pr(N) (the silent allele probability) and the dotted line represents the probability of dropout when amplification efficiency is not taken into account.

probability of dropout, $Pr(O^l_{ar} < Z)$ we consider the combination of amplification efficiency and peak height variability that lead to potential dropout. This is achieved by:

$$Pr(O^l_{ar} < Z) = \int_{A^l} Pr(O^l_{ar} < Z|E^*, A^l) Pr(A^l) \mathrm{d}A^l \quad (8)$$

where $E^l_{ar} = E^* \times A^l$ and

$$Pr(O^l_{ar} < Z|E^*, A^l) = \int_{O=0}^{Z} Pr(O) Pr\left[\log\left(\frac{O}{E^l_{ar}}\right)\right] \mathrm{d}O$$

In words, amplification efficiencies are considered that would change the expected height from E^* to E^l_{ar} and then the probability of dropout is considered for a peak at height E^l_{ar}, integrated across all values of amplification efficiency. Using these models the probability given by Eq. (8) is dominated by peak height variability at low expected heights and by amplification efficiencies at high peak heights. Fig. 2 shows the results of applying equation 8 to peak heights E^l_{ar} from Z (30 rfu) to 1000 rfu and using $\sigma^2 = 0.03$. The model in Eq. (8) relies on having an expected peak height of E^l_{ar}, and this can be approximated from the average allelic peak height across a profile, or by application of a model that accounts for degradation and molecular weight. Also shown in Fig. 2 is the probability of dropout (as seen in Eq. (8)) without taking into account locus amplification efficiency.

From the results seen in Fig. 2 an approximate value that could be used for distinguishing dropout from a null allele could be approximately 350 rfu. The probability of dropout line in Fig. 2 (using Eq. (8)) is that which would be used to consider the probability of dropout if there were multiple allelic peaks present in a profile. For example a 1:1 mixture where only a single peak is present could be treated similarly to an autosomal profile when deciding whether the risk that not all information has amplified above the analytical threshold (AT) is negligible. A probability cutoff value could be chosen, such as 0.001, and the corresponding height used as a threshold (from the dotted line in Fig. 2 this would be approximately 200 rfu).

2.2. Interpreting a major component

We will use the results of the probabilistic estimation of peak height variability from Section 1.2 to address mixture proportion guidelines.

Consider first a two person Y-STR profile, with L loci. The individuals have contributed DNA in unequal amounts. We define the proportion of contributor 1 as m_1 and of contributor 2 as m_2. These are unknown. To develop guidelines, which allow us to assign the apparent major component to the major contributor with some level of confidence we need to consider:

- Varied mixture proportions.
- Varied profile intensities.
- What probability do we accept for the incorrect assignment of a major contributor.

There are L-d loci showing two peaks. d loci show one peak and that is either because both peaks are the same allele, or one allele has dropped, or one allele is silent.

Consider initially the loci showing two peaks. If we call the larger peak at each two allele locus E_M and the smaller E_m, then we could imagine that all L-d of the larger peaks are attributable to the major and the remainder to the minor. Term this arrangement S_1. However there are many other potential arrangements. Consider initially the swap of two peaks at one specific locus such that the smaller peak is assigned to the major and the larger to the minor. Term this arrangement S_2.

The probability of any particular assignment, j, is estimated as $\int_M \prod_l \prod_a \prod_r Pr(O^l_{ar}|S^l_j, M) Pr(M) \mathrm{d}M$. We require

$$\frac{\int_M \prod_l \prod_a \prod_r Pr(O^l_{ar}|S^l_1, M) Pr(M) \mathrm{d}M}{\int_M \prod_l \prod_a \prod_r Pr(O^l_{ar}|S^l_2, M) Pr(M) \mathrm{d}M} > \alpha$$

where α is a likelihood ratio threshold we wish to apply. The formulation above avoids assuming a known ratio, but is complex to apply. We again make a simplifying assumption that the profile ratio is known and available to be applied to the methodology below. In practise this is likely to be average ratio observed across the profile. There will be some increased uncertainty associated with this practise.

We start with a mixture that possesses contributors in proportions 2/3 and 1/3 (or 2:1 in ratio). Two peaks are observed at O_1 and O_2 where $O_1 \sim 2 \times O_2$. There is a possibility that O_1 has been donated by the major contributor and O_2 by the minor. There is also possibility that O_1 has been donated by the minor contributor and O_2 by the major, but peak height variability has caused the apparent flip in relative peak heights. Possibility 1 is going to be more likely than possibility two, as it is a better fit with the mixture proportions, but the question for interpretation is whether the alternative is so unlikely that it can be discounted. This will depend on the absolute peak heights of O_1 and O_2.

If the profile is adhering to the mixture proportions then the expected heights for the major (E_1) and minor (E_2) are:

$$E_1 = \frac{2}{3}(O_1 + O_2) \text{ and } E_2 = \frac{1}{3}(O_1 + O_2)$$

Therefore the probability of the observed data if O_1 is from the major and O_2 is from the minor (denoted as haplotype set 1, S_1) is:

$$\begin{aligned} p(O|E, S_1) &= Pr\left[\log\left(\frac{O_1}{E_1}\right)\right] \times Pr\left[\log\left(\frac{O_2}{E_2}\right)\right] \\ &= Pr\left[\log\left(\frac{O_1}{\frac{2}{3}(O_1 + O_2)}\right)\right] \times Pr\left[\log\left(\frac{O_2}{\frac{1}{3}(O_1 + O_2)}\right)\right] \end{aligned}$$

where the probabilities are enumerated using Eq. (3a).

And the alternative where O_2 if from the major and O_1 is from the minor (denoted as haplotype set 2, S_2) is:

$$p(O|E, S_2) = Pr\left[\log\left(\frac{O_2}{\frac{2}{3}(O_1 + O_2)}\right)\right] \times Pr\left[\log\left(\frac{O_1}{\frac{1}{3}(O_1 + O_2)}\right)\right]$$

Whether or not we are willing to interpret O_1 as the major then is determined by the peak heights and the confidence we desire in the assignment. Let us say our criterion was: $\frac{p(O|E,S_1)}{p(O|E,S_2)} > \alpha$. This criterion requires the evidence to be α times more likely under S_1 than under S_2. This should not be confused with stating that S_1 is α times more likely than S_2.

Fig. 3 shows the value of $\frac{p(O|E,S_1)}{p(O|E,S_2)}$ over a range of heights for O_1 and O_2, where we consider a 2:1 mixture and assume $O_1 = 2O_2$, a 3:1 mixture and assume $O_1 = 3O_2$, a 4:1 mixture and assume $O_1 = 4O_2$ and a 5:1 mixture and assume $O_1 = 5O_2$.

From Fig. 3 the thresholds on the height of the largest peak that are suggested for the assignment of the largest peak as coming from the major component for 2, 3, 4 or 5 to 1 mixtures (where both peaks have been observed in approximately that configuration) are 1022, 455, 304 and 234 rfu, respectively using a threshold of $\alpha = 1000$.

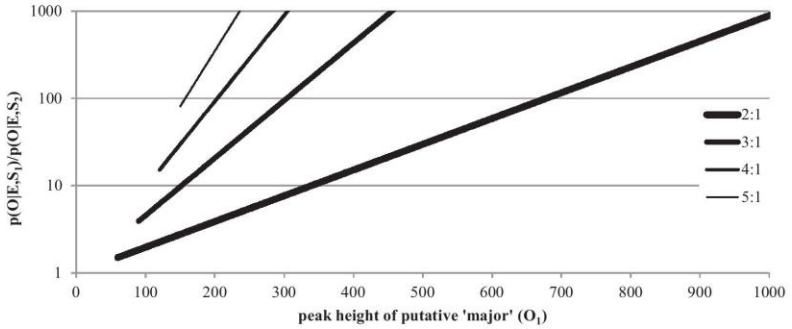


Fig. 3. Likelihood ratio considering the probability of the observed peaks if the higher peak is from the major contributor as opposed to the smaller peak being from the major contributor.

We can also consider the situation where only a single peak has been observed at a locus and we wish to know when it can be assigned to the major with confidence. Again a formal treatment would see the probability of the observed data considered in light of the two competing scenarios:

S_1: the major contributor is the source of the observed peak,

S_2: the minor is the source of the observed peak with the major having dropped out, and integrated across all mass parameters so that assumptions regarding mixture proportions, DNA amount, degradation and amplification efficiency are not required. Again we make some simplifying assumptions in order to simply determine a threshold for this data. If the single peak is height O and the mix proportions are m_1 and m_2 then we evaluate the described scenario by:

$$p(O|E,S_1) = \frac{1}{2}Pr\left[\log\left(\frac{O_1}{O_1}\right)\right]Pr(O^I_{ar} < Z|E = \frac{m_2}{m_1}O_1) + \frac{1}{2}Pr\left[\log\left(\frac{O_1}{O_1}\right)\right] \quad \text{and}$$

$$p(O|E,S_2) = Pr(O^I_{ar} < Z|E = \frac{m_1}{m_2}O_1)Pr\left[\log\left(\frac{O_1}{O_1}\right)\right]$$

Note that under $p(O|E,S_1)$ we consider that either the minor has dropped out (left part of the equation) or that it is masked under the major (right part of the equation), each with equal prior probability. Notice here that we simplify to the point that we are not considering the impact that amplification efficiency may be having on the locus, however this is likely to lead to a conservative threshold (i.e. requiring higher fluorescence in order to interpret) and so is justifiable. Using the above formulations the points at which $\frac{p(O|E,S_1)}{p(O|E,S_2)} > 1000$ is obtained for 2:1, 3:1 or 5:1 are 93, 60 or 40 rfu, respectively. The threshold of interpretation can be plotted against the mixture ratios in order to obtain a relationship between the two properties that allows thresholds to be considered for a range of ratio. This can be seen in Fig. 4, where a power trend line has been fitted to the data.

A similar approach can be used to consider more complex mixtures just as long as all genotypic sets are considered that include, dropout or masking when determining the likelihood ratio. Without deriving in the same detail as above, we provide some scenarios in Table 2 as examples for 3 person profiles. Note that the first row considers the same imbalances as the two person scenario above and so the thresholds are the same.

Although the same system of threshold determination could be generated for profiles of any complexity, we suggest that the higher order mixtures (for example greater than three males, or profiles complicated in some other way such as differential degradation or inhibition) may be out of the scope of manual interpretation and better suited to analysis in probabilistic systems.

2.3. Interpreting a minor component

We consider here the situation where a single peak has been observed at a locus in a profile where the two contributors are in uneven proportions. Interpreting a minor component can be treated in the same way as a major component, although the added complexity that the minor peak could be masked under the major

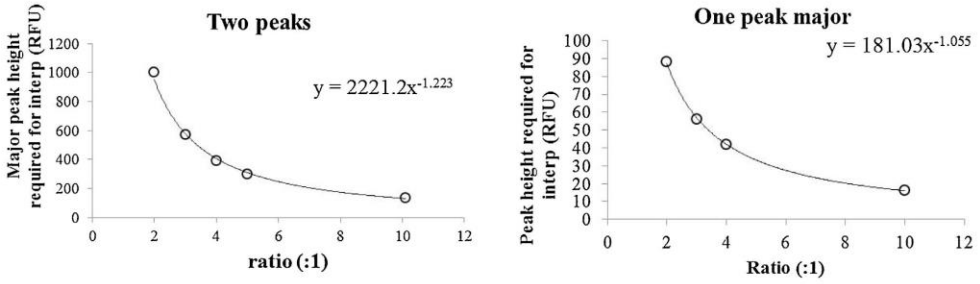


Fig. 4. Relationships between the mixture ratio for two person mixtures and the threshold required for the higher peak to be assigned to a major component.

Table 2
DNA profile examples and potential thresholds for assigning the larger peak to a major component of a 3 person mixture.

DNA Profile description	Scenarios	Value of x when $\frac{p(O\|E,S_1)}{p(O\|E,S_2)} > 1000$
3 person profile with 3 peaks	S_1: major = [A], minors = [B] and [C]	x:y = 2:1—1022 rfu
$O_A = x$ rfu	S_2: major = [B] or [C], minors = [A] and [C] or [B]	x:y = 3:1—455 rfu
$O_B = y$ rfu		x:y = 4:1—304 rfu
$O_C = y$ rfu		x:y = 5:1—234 rfu
$x > y$		
3 person profile with 2 peaks	S_1: major = [A], minors = [A] and/or [B] and/or [Q]	x:y = 2:1—4500 rfu
$O_A = x$ rfu	S_2: major = [B] or [Q], minors = [A] and [Q] and/or [B]	x:y = 3:1—738 rfu
$O_B = y$ rfu		x:y = 4:1—389 rfu
$x > y$		x:y = 5:1—274 rfu
3 person profile with 1 peak	S_1: major = [A], minors = [A] and/or [Q]	x:y = 1:1—453 rfu
$O_A = x$ rfu	S_2: major = [Q], minors = [A] and possibly [Q]	x:y = 2:1—286 rfu
		x:y = 3:1—227 rfu
		x:y = 4:1—199 rfu
		x:y = 5:1—182 rfu

must be taken into account. There are two scenarios that we consider:

1) S_1: the peak at height O is made up of two individuals donating an A allele and its observed height is equal roughly to what is expected. The numerator of the LR may be approximated by the density of the normal distribution $N(0, c^2/\mathrm{O})$ using the mode of the distribution for the variance
2) S_2: the peak at height O is only from the major contributor and the other peak, expected to have occurred as height $O\frac{m_2}{m_1}$ has dropped out. This probability can be determined by Eq. (8).

Given an observed peak height O for allele A, we must consider the point where the ratio of likelihoods $\frac{p(O|E,S_1)}{p(O|E,S_2)}$ exceeds some value (α). To be consistent with previous interpretations we will use a value of $\alpha = 1000$. Define the proportion that the major contributor is donating to the mixture as m_1 and the proportion from the minor contributor is m_2 (where for a two person mixture $m_1 = 1 - m_2$). In this simplified treatment of the data we are not going to consider the effects that amplification efficiency may have on the interpretation and so are losing some power.

From Fig. 5, and using an LR threshold of 1000 (arbitrarily chosen) it would be acceptable to designate the minor peak in a mixture as masked by the major for 2, 3, 4 and 10 to 1 mixtures when the height of the single observed peak was 270, 520, 1000 and 2340 rfu, respectively. Again, the threshold of peak height required for interpretation of a masked minor component can be plotted against the mixture ratio as in Fig. 6.

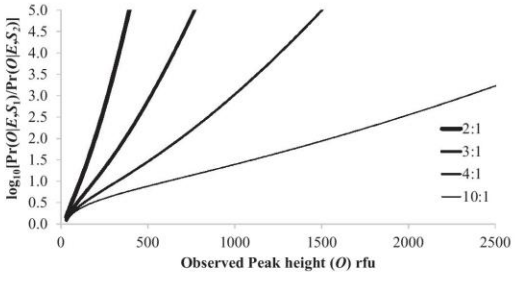


Fig. 5. The probability of a single observed peak given the scenario that the minor is masked under the major peak vs that the minor peak has dropped out for various mixture proportions.

2.4. Variation in mixture proportions across a profile

The thresholds given above are developed using simplifying assumptions. It is likely that there will be variability in the mixture proportions across a profile and the question may be asked as to whether the mixture proportion at any one locus deviates sufficiently from a profile average that it casts doubt on the interpretation, particularly for higher order mixtures. Consider a locus, which has mixture proportions of m_1 and m_2 and a total observed height (the sum of the peaks) of O_T rfu. Peaks are expected to be present at $E_1 = O_T m_1$ and $E_2 = O_T m_2$ rfu. We know from Eq. 3a that $O_T m_1$ will be observed at height O_1 for some proportion of the time as given by equation:

$$O_1 = O_T m_1 \times 10^x$$

where x is determined as the value of $\log(O_1/E_1)$ for which a normal distribution $N(0, c^2/E_1)$ is at the desired quantile. If we consider both peaks together (and make the simplifying assumption that the imbalance probability is apportioned evenly between the two peaks) then we are interested in the point at which both values of $\log(O/E)$, considered together, would reach the desired quantile. To demonstrate this concept we have considered mixtures with ratios of 1:1, 2:1 and 10:1. Fig. 7 shows the 95% intervals on the mixture proportion variation that would be expected for a 1:1 mixture.

Typically, interpretation thresholds that deal with allowable variation in the relative contribution of contributors are based on mixture proportions. Threshold such as $|D| \leq 0.2$ when average peak height is greater than 300 rfu have been suggested for autosomal profiling systems [16]. That is, the mixture proportion of any individual locus being interpreted must fall within ± 0.2 of the

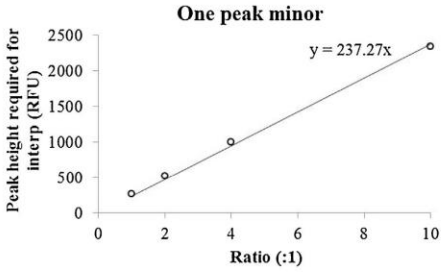


Fig. 6. Relationships between the mixture ratio for two person mixtures and the threshold required for assigning a minor allele as masked by the major.

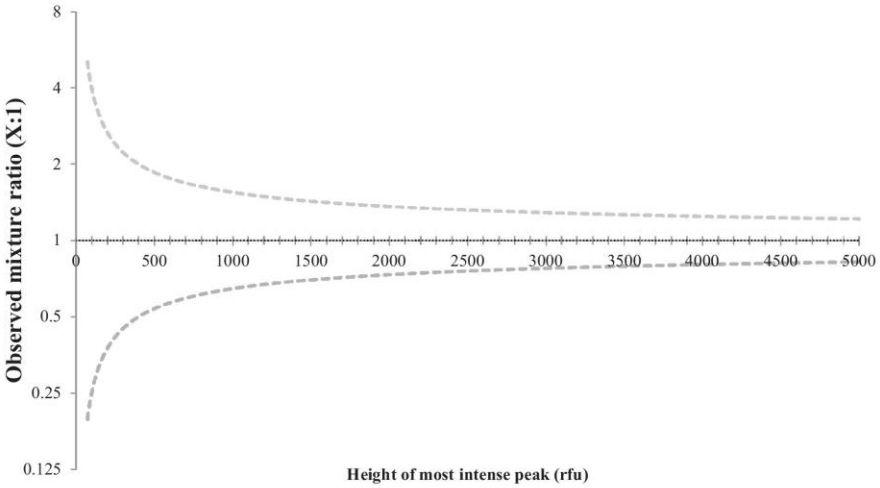


Fig. 7. Variation expected in 1:1 mixture ratio for a range of peak intensities. Dashed lines show the 95% interval bounds.

profile-wide average value. We demonstrate the performance of a mixture proportion (Mx) threshold of ±0.2 on data that has mixture ratios of 1:1, 2:1 and 10:1 (or mixture proportions of 0.5, 0.34, 0.09) in Fig. 8.

Note that in Fig. 8 for the 10:1 profile there is no lower bound on the acceptable mixture proportion of the minor contributor. Also from all three graphs in Fig. 8 it appears that the mixture proportion threshold of ± 0.2 covers most data that is above a typically 'stochastic range'. Another method that could be used would be to determine at which peak intensity point the mixture proportion reached 0.5, which for the 2:1 and 10:1 ratios would be 580 rfu and 164 rfu respectively. Whilst this second method may allow more interpretation due to its increased flexibility it is also more difficult to apply.

2.5. Determining peak height variability without continuous analysis software

If a laboratory does not have access to computing systems that can determine the peak height variability in the manner described by Section 1.2 then an approximate value will need to be determined using empirical observations. This can be done by comparing the observed peak heights with the expected peak heights (determined from the stutter peak heights and the stutter ratio for that allele and locus combination) and adjusting the value of c^2 until quantile bounds have their expected coverage. This is basically retrofitting the variance constant to empirical observations as seen in Fig. 1. This can be done by:

(a) Determine the expected height for the allelic peak at locus l in profile i, E_i^l, by $E_{ar}^l = \frac{O_{(a-1)r}^l}{\overline{\pi}_a^l}$

(b) Graph $\log\left(\frac{O_{ar}^l}{E_{ar}^l}\right)$ against E_{ar}^l for allelic peaks

(c) Graph the 95% upper bound by $+Z\sqrt{\frac{2c^2}{E}}$, where Z determines the quantile

(d) Adjust c^2 until only 2.5% of the data falls above the upper 95% bound

Doing the above for the dataset described in Section 1.2 gives a value of $c^2 = 5.4$, which is in the mode of the variance constant distribution. This is lower than the value of $c^2 = 9$, which is the mean of the distribution that was used for thresholds. This

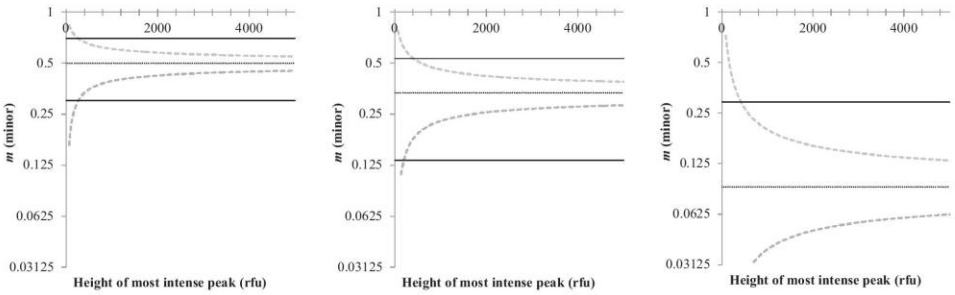


Fig. 8. Variation expected in 0.5 (left), 0.34 (middle) and 0.09 (right) mixture proportions, m, for a range of peak intensities. Data is shown only for the minor contributor mixture proportion. Dashed lines show the 95% interval bounds. Solid lines show a value of ±0.2 from the known mixture proportion.

difference is likely to decrease as the dataset used to derive it increases in size, at which point the manually derived value is likely to converge to the continuously determined value.

The second aspect required for interpretation is an inter-locus variance, σ^2, for the amplification efficiency model:

$$\log_{10}\left(A^l\right) \sim N(0, \sigma^2) \quad (9)$$

Determination of a value for σ^2 can be carried out by:

(a) Determining the average peak height for each of the I profiles, $\overline{E_i}$
(b) Calculate the amplification efficiency factor for each locus that would be required for it to equal $\overline{E_i}$. For example if the observed height of the allele at locus l in profile i is O_i^l, then the amplification efficiency is calculated by $A_i^l = \frac{O_i^l}{\overline{E_i}}$
(c) The likelihood of the dataset, L, is calculated by $L = \prod_i \prod_l \Pr\left(A_i^l\right)$, where $\Pr\left(A_i^l\right)$ is modelled by Eq. (9)
(d) The value of σ^2 is varied to maximise L.

Carrying out this procedure for the dataset described in Section 1.2 yields an amplification efficiency variance value of $\sigma^2 = 0.04$, which is close to the value given in Section 1.2 of $\sigma^2 = 0.03$. The increased value for σ^2 when using the method described above is likely due to the fact that in the manual method above there is no degradation term, therefore any differences in peak heights will need to be accounted for by amplification efficiency. The continuous model described in Section 1.2 does have a model for degradation and hence the lower amplification efficiency variance. The approximation of no degradation using the manual method is likely to have a small impact on the value of σ^2 as long as the profiles being considered do not exhibit marked degradation.

2.6. Assessing the interpretation models

In order to assess the performance of the suggested continuous and binary interpretation models we generated mixed DNA profiles from two, three and four male individuals in dilutions as specified in Table 3. Mixed DNA samples were amplified in triplicate as per manufacturer's instructions to obtain a total of 90 Yfiler Plus profiles.

2.7. Probabilistic models

The method described in Ref. [1] and Section 1 of this paper described the biological and statistic models used to analyse Y-STR data in a fully continuous manner. Section 2 uses these continuous models in order to develop interpretational thresholds than can be applied in a manual fashion, or with the assistance of an Excel spreadsheet in a relatively straightforward manner. Before applying the thresholds to mixture data we must first show that the continuous models on which they are based are indeed functioning correctly. We apply the continuous methodology using a modified version of STRmix™ (http://strmix.esr.cri.nz/) [1] (which we refer to here as STRmixY, and is currently in a developmental state) to carry out this function. It is impractical to display all results for all mixtures analysed in this study and so in Table 4 we show the results for the 1:1 mixture deconvolution, displaying all haplotype combinations and weights generated by the continuous method. The mixture, when generated, was in reality slightly divergent from the target 1:1 and the continuous interpretation picked this up, hence the weights in Table 4 are not split evenly between contributor 1 and 2. A copy of the profile is given in Fig. 9 and this divergence can be seen. In Table 4 the genotype sets associated with the known contributors are bolded. It can be seen in all but one instance the correct genotype set was given the highest weight. In the one instance that the highest weight was not given to the known genotypes set (in DYS448), the weights correctly reflect the results obtained in the epg (Fig. 9).

Fig. 10 shows the target mixture proportions of the analysed profiles and the mean of the posterior distribution for mixture proportions obtained from deconvolution by STRmixY.

From these analyses the deconvolution of mixed Y-STR data can proceed in the same fashion as for autosomal data. There are other aspects of the continuous deconvolution of these samples, such as the assessment of locus amplification efficiencies, allelic and stutter peak height variability, degradation levels and a number of analysis diagnostics that we do not show here for reasons of

Table 3
Mixtures constructed for analysis.

Tubes	Mixture proportions for contributor				Total DNA added to PCR (pg)
	One	Two	Three	Four	
1–4	0.50	0.50			500,200,50,20
5–8	0.33	0.67			
9–12	0.17	0.83			
13–15	0.09	0.91			
16–18	0.02	0.98			
19–21	0.33	0.33	0.33		500,200,50
22–24	0.50	0.33	0.17		
25–27	0.25	0.25	0.25	0.25	
28–30	0.40	0.30	0.20	0.10	

Table 4
Weights for 500 pg 1:1 mixture PCR 1. Contributor 1 and 2 percentages are the posterior mean mixture proportions from STRmixY. Bolded values indicate the correct genotypic assignments.

Locus	Contributor 1 Genotype (59%)	Contributor 2 Genotype (41%)	Weight
DYS576	**[16]**	**[17]**	**0.914**
	[17]	[16]	0.086
DYS389I	**[12]**	**[12]**	**1.000**
DYS635	**[21]**	**[21]**	**1.000**
DYS389II	**[29]**	**[28]**	**0.906**
	[28]	[29]	0.094
DYS627	**[19]**	**[19]**	**1.000**
DYS460	**[11]**	**[10]**	**0.926**
	[10]	[11]	0.074
DYS458	**[15]**	**[15]**	**1.000**
DYS19	**[14]**	**[14]**	**1.000**
YGATAH4	**[10]**	**[11]**	**0.923**
	[11]	[10]	0.077
DYS448	[19]	[20]	0.753
	[20]	**[19]**	**0.247**
DYS391	**[10]**	**[10]**	**1.000**
DYS456	**[15]**	**[14]**	**0.904**
	[14]	[15]	0.096
DYS390	**[23]**	**[22]**	**0.916**
	[22]	[23]	0.084
DYS438	**[10]**	**[10]**	**1.000**
DYS392	**[11]**	**[11]**	**1.000**
DYS518	**[40]**	**[39]**	**0.634**
	[39]	[40]	0.366
DYS570	**[18]**	**[20]**	**0.926**
	[20]	[18]	0.074
DYS437	**[16]**	**[16]**	**1.000**
DYS449	**[26]**	**[28]**	**0.855**
	[28]	[26]	0.145
DYS393	**[13]**	**[13]**	**1.000**
DYS439	**[11]**	**[11]**	**1.000**
DYS481	**[27]**	**[25]**	**0.922**
	[25]	[27]	0.078
DYF387S1	**[38]**	**[37]**	**0.917**
	[37]	[38]	0.083
DYS533	**[11]**	**[11]**	**1.000**

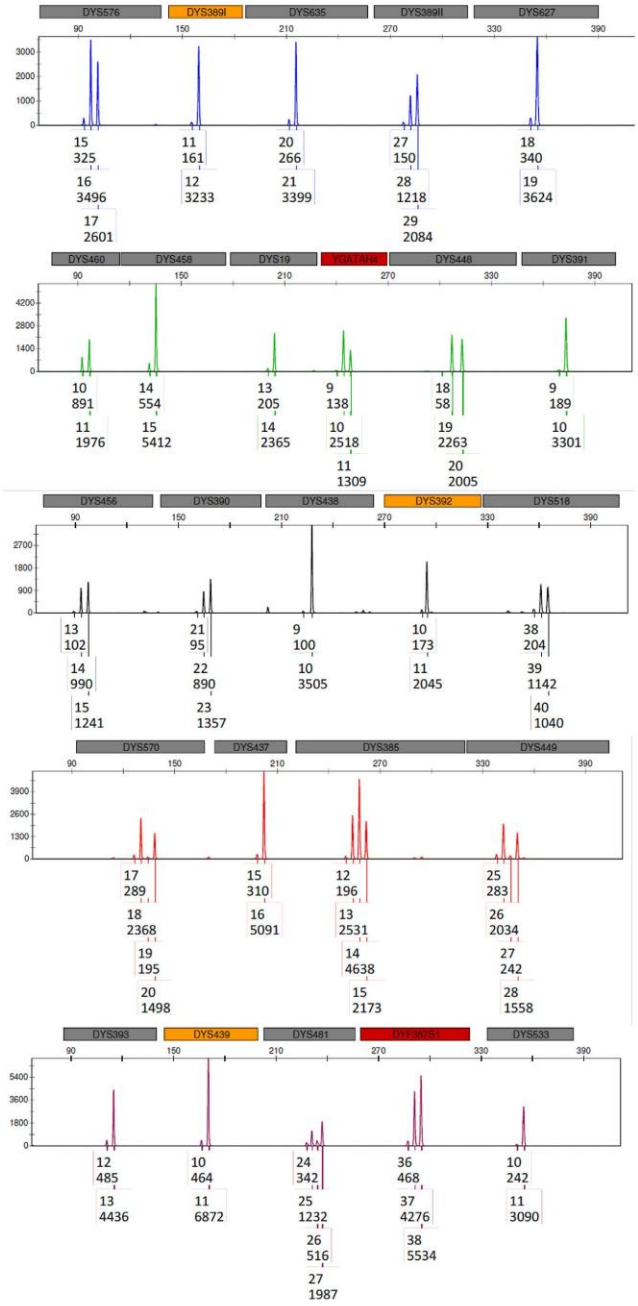


Fig. 9. epg for 500 pg 1:1 mixture PCR 1.

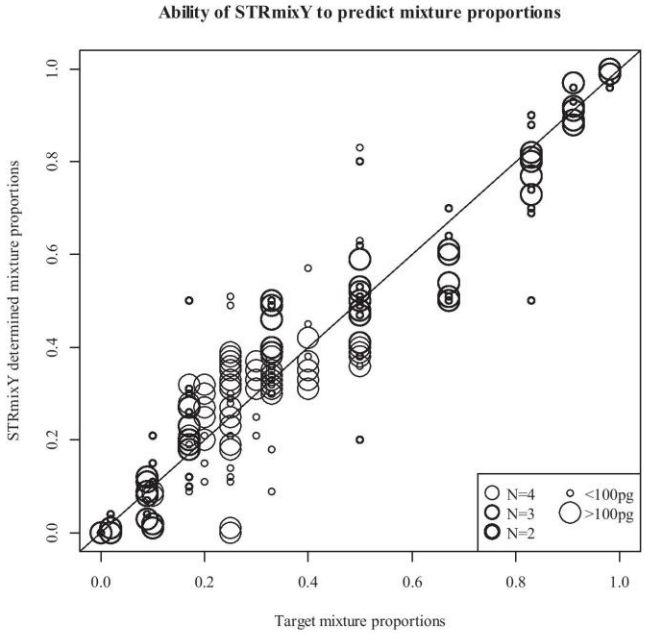


Fig. 10. Target mixture proportions compared with posterior mean mixture proportions from deconvolution.

brevity, but give further assurances that the continuous methods and underlying models are describing the observed data well. From the results of the probabilistic modelling we are confident to proceed on to testing the manual interpretation thresholds based on these models.

2.8. Threshold models

We now subject the two person profiles detailed in Table 3 to the threshold based interpretation rules generated in Section 2. We generate the average mixture ratio, $\overline{M_r}$, by

1. Averaging value of the log of highest observed peak divided by the lowest observed peak at each locus to provide the apparent mixture ratio at a locus, $\hat{M}_r^l$, and
2. Taking 10 to the power of this value $\overline{M_r} = 10^{\frac{1}{L}\sum_l \log(\hat{M}_r^l)}$. This can then be converted to an average mixture proportion by, $\overline{Mx} = \frac{\overline{M_r}}{1+\overline{M_r}}$.

We then apply rules rigidly using Excel to all two person mixtures. For the major component at a locus to be interpretable:
If there are two observed peaks:

1. The observed mixture ratio must be within ±0.2 of $\overline{Mx}$ (as per Section 2.4) (in terms of $\hat{M}_r^l$, $\frac{\overline{Mx}-0.2}{1.2-\overline{Mx}} \leq \hat{M}_r^l \leq \frac{\overline{Mx}+0.2}{0.8-\overline{Mx}}$) and
2. The peak height of the major peak must be greater than or equal to $2221.2 \times \overline{M_r}^{-1.223}$ (as per Fig. 4 left pane).

If there is one peak:

3. Then it must be greater than or equal to $181.03 \times \overline{M_r}^{-1.055}$ (as per Fig. 4 right pane).

Even given these three rules there is a requirement for an additional rule that prevents interpretation of a major (or minor) component in a mixture when the observed peak heights are too close. We create this rule with consideration of the created and observed ratios for the two person mixtures at different concentrations, as seen in Fig. 11.

Fig. 11 shows that there is overlap between individual locus mixture ratios at all levels of input DNA and overlap of profile averages when 50 pg or less is added to the PCR. As expected the observed ratio for the more disparately prepared mixtures is much less than the target ratios. This is because the expected peak heights of the minor components are below the analytical threshold, and so all observed values will be those that are stochastically higher than expected. Given the results of Fig. 11 we implement one additional interpretation threshold for interpreting major components:

4. For a locus with two peaks the observed mixture ratio at that locus must be greater than 2:1, i.e. $\hat{M}_r^l \geq 2$.

Note that rule 4 is somewhat more arbitrarily chosen than the other interpretational thresholds. The modelling of data to develop this threshold is more complex because it involves the observed average mixture ratio, the mixture ratio at the locus being examined and the height of the peak involved. Multiple regression could be performed to develop this rule, however we feel that it is

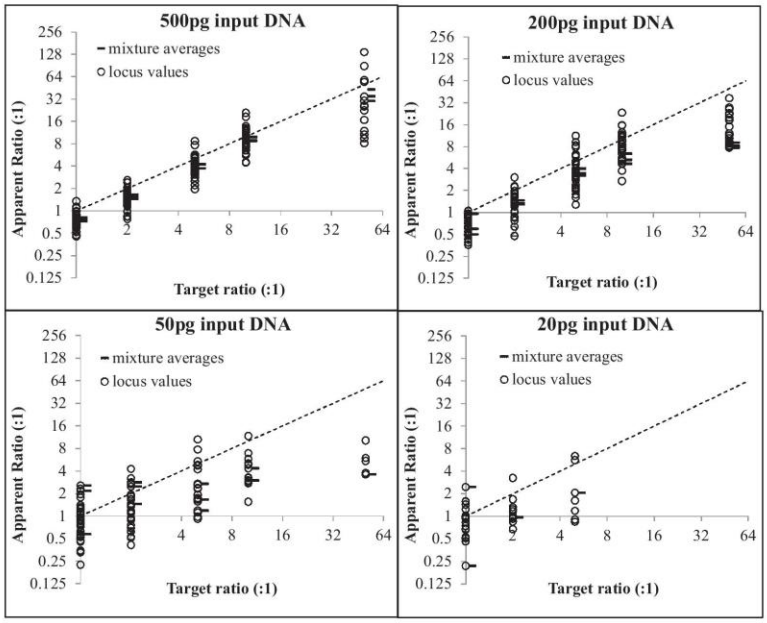


Fig. 11. Target and apparent mixture ratios for two person mixtures at different total DNA amounts. Profile averages are given as black bars and individual locus values are given as hollow circles.

likely to become overly cumbersome to apply in practise and so we simply choose a value that should perform conservatively.

To interpret a minor component, thresholds 1, 2 and 4 apply, and the modified rule 3 becomes:

5. If there is one peak present then it must be greater than or equal to $237.27 \times \overline{M_r}$ (as per Fig. 6).

The results of applying these rules to the two person profiles on all non-duplicated loci (i.e. all loci excluding DYS385) are shown in Table 5. The maximum number of loci at which a component could be interpreted is therefore 24. The number of loci that can potentially be interpreted will depend on the ratio (i.e. only single peak loci can be interpreted from 1:1 mixtures) and the number of loci where information is detected.

For the results in Table 5 all interpreted alleles for both major and minor contributors were assigned correctly. The sample asterisked in Table 5, possessed a genotypic assignment for major and minor at a locus where the alleles of major and minor contributors were different from each other. For the locus in question the observed average mixture ratio was 1.5:1 and the locus at which the interpretation was made possessed two alleles; 10 at 891 rfu and 11 at 1976 rfu (locus DYS460 from Fig. 9). Applying the thresholds 1–5 interpreted a major contributor at this locus as having an 11 allele. All other loci interpreted for this profile contained only a single peak (and so both contributors could be interpreted as having this allele). While the above interpretation is correct, in that one of the contributors has this allele, we highlight it as the target mixture proportion for this profile was 1:1 and so we expected that no component would have been interpreted at this locus. It is worth noting that the highest weight given by the continuous interpretation method to this sample (Table 4) for any of the loci with two distinct alleles was 0.93, at locus DYS460 and DYS570. In total 756 assignments were made of a major contributor and 399 assignments of a minor contributor out of the (54×24) 1296 potential assignments.

In a broader sense the results of the manual interpretation give reassurance that the continuous models used to describe DNA profile behaviour can be manipulated to generate binary thresholds for use in manual profile interpretations. It is also worth noting that the dataset to which these thresholds have been applied is not the training set used to generate the thresholds in the first place. This has been done so as to evaluate their use as they will be applied in casework.

Table 5
The number of loci at which the major and minor contributor can be assigned for the two person mixtures outlined in Table 3 (results of 3 PCRs combined). The asterisked results are explained within the text.

Ratio	DNA (pg)	Major loci interpreted	Minor loci interpreted	Loci with alleles detected
1:1	500	**37***	**37***	72
2:1		38	38	72
5:1		70	70	72
10:1		72	63	72
50:1		72	16	72
1:1	200	36	36	72
2:1		36	36	72
5:1		63	49	72
10:1		70	24	72
50:1		72	20	72
1:1	50	32	3	72
2:1		33	4	72
5:1		36	0	72
10:1		36	1	71
50:1		28	2	72
1:1	20	9	0	62
2:1		3	0	54
5:1		13	0	52

3. Conclusion

Y-STR data can be crucial to forensic investigations, particularly when autosomal STR information has failed to yield an informative result. The relatively low proportion of cases that utilise Y-STRs, combined with difficulties in evaluating an *LR*, due to their mode of inheritance, has meant that the advances seen in autosomal STR analysis have not been mirrored for Y-STR data. This leaves many laboratories with limited power to develop interpretation guidelines and calculate evidential weights for Y-STR data, particularly for mixed profiles. In this work we develop methods that overcome some of these limitations by suggesting ways that laboratories can develop interpretational guidelines, based on the same probabilistic theory that has advanced autosomal interpretation. We demonstrate the generation of interpretation guidelines based on dropout, mixture proportion, peak balances and major/minor component configuration and applied them to laboratory generated mixtures. The reason we demonstrate methods for generating interpretational guidelines at all is due to the realisation that a publically available probabilistic system for Y-STR data interpretation is unlikely in the short term. In the interim some methods for manual interpretation are still required.

We also applied the probabilistic methods directly (without the use of thresholds) to the same data. As has been demonstrated in a number of works (for example see [1]) the application of thresholds (which are typically designed to be conservative) is wasteful of data and leads to only a fraction of the generated mixed DNA profiles able to be used for calculation of an *LR* .

We note that we have not considered the effects of multiple PCR replicates on the ability to manually interpret Y-STR mixtures in this work. The use of multiple replicates will increase the confidence of the analyst in their interpretation and hence could be used to reduce the interpretational requirements (i.e. lower peak heights for interpreting a major component).

Conflict of interest

The authors of this work are the technical developers of the commercially available software product STRmix™. The developers in no way profit financially from commercial STRmix™ activities.

Acknowledgements

This work was supported in part by grant 2014-DN-BX-K028 from the US National Institute of Justice. Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the U.S. Department of Justice or Commerce. Certain commercial equipment, instruments, or materials (or suppliers, or software) are identified in this paper to foster understanding. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

References

[1] D. Taylor, J.-A. Bright, J. Buckleton, The interpretation of single source and mixed DNA profiles, Forensic Sci. Int.: Genet. 7 (2013) 516–528.
[2] H. Haned, Forensim: an open-source initiative for the evaluation of statistical methods in forensic genetics, Forensic Sci. Int.: Genet. 5 (2011) 265–268.
[3] K. Lohmueller, N. Rudin, Calculating the weight of evidence in low-template forensic DNA casework, J. Forensic Sci. 58 (2013) 234–259.
[4] C.D. Steele, D.J. Balding, Statistical evaluation of forensic DNA profile evidence, Annu. Rev. Stat. Appl. 1 (2014) 361–384.
[5] R. Puch-Solis, L. Rodgers, A. Mazumder, S. Pope, I. Evett, J. Curran, et al., Evaluating forensic DNA profiles using peak heights, allowing for multiple donors, allelic dropout and stutters, Forensic Sci. Int.: Genet. 7 (2013) 555–563.
[6] D.J. Balding, J. Buckleton, Interpreting low template DNA profiles, Forensic Sci. Int.: Genet. 4 (2009) 1–10.
[7] M.W. Perlin, M.M. Legler, C.E. Spencer, J.L. Smith, W.P. Allan, J.L. Belrose, et al., Validating TrueAllele® DNA mixture interpretation, J. Forensic Sci. 56 (2011) 1430–1447.
[8] J.-A. Bright, J. Curran, J. Buckleton, Modelling PowerPlex® Y stutter and artefacts, Forensic Sci. Int.: Genet. 11 (2014) 126–136.
[9] M.M. Andersen, J. Olofsson, H.S. Mogensen, P.S. Eriksen, N. Morling, Estimating stutter rates for Y-STR alleles, Forensic Sci. Int.: Genet. Suppl. Ser. 3 (2011) e192–e193.
[10] D. Taylor, J.-A. Bright, C. McGoven, C. Hefford, T. Kalafut, J. Buckleton, Validating multiplexes for use in conjunction with modern interpretation strategies, Forensic Sci. Int.: Genet. 20 (2015) 6–19, doi:http://dx.doi.org/10.1016/j.fsigen.2015.09.011.
[11] J.-A. Bright, D. Taylor, J.M. Curran, J.S. Buckleton, Developing allelic and stutter peak height models for a continuous method of DNA interpretation, Forensic Sci. Int.: Genet. 7 (2013) 296–304.
[12] P. Gill, J. Curran, K. Elliot, A graphical simulation model of the entire DNA process associated with the analysis of short tandem repeat loci, Nucleic Acids Res. 33 (2005) 632–643.
[13] D. Taylor, J. Buckleton, J.-A. Bright, Factors affecting peak height variability for short tandem repeat data, Forensic Sci. Int.: Genet. 21 (2016) 126–133.
[14] J.M. Butler, Chapter 7—Low-level DNA and complex mixtures, in: J.M. Butler (Ed.), Advanced Topics in Forensic DNA Typing: Interpretation, Academic Press, San Diego, 2015, pp. 159–182.
[15] M.M. Andersen, H.S. Mogensen, P.S. Eriksen, J.K. Olofsson, M. Asplund, N. Morling, Estimating Y-STR allelic drop-out rates and adjusting for interlocus balances, Forensic Sci. Int.: Genet. 7 (2013) 327–336.
[16] J.-A. Bright, J. Turkington, J. Buckleton, Examination of the variability in mixed DNA profile parameters for the Identifiler™ multiplex, Forensic Sci. Int.: Genet. 4 (2010) 111–114.

7.2 A variable number of contributors

As profiles become larger in the number of regions targeted and more sensitive, the forensic community faces an ever-increasing problem of complexity. As complexity increases, it becomes more and more onerous to assign a number of contributors to the profile. In the context of assigning a probability to the observed data, given a defence or prosecution proposition, the number of contributors is one of the few (if not the last) nuisance parameter that users must still place all their belief in a single value derived through manual interpretation. With new technology on the horizon (called massively parallel sequencing, which is briefly explained later in the thesis) the number of loci looks set to increase further, and be supplemented with the underlying DNA sequences, adding yet further complexity. The forensic biology community will soon exist in a world where the data is simply too complex for a human pre-assessment on any nuisance parameters to meaningfully occur. Some argue that this point has already arrived. A method is required to treat the number of contributors to a DNA profile as a nuisance parameter that can be integrated over within the LR model. The development of treating the number of contributors as a nuisance parameter was the theoretical drive for the work published in this section of the thesis. There was also a very clear practical drive for this work. As the use of STRmix™ increased around Australia and New Zealand laboratories during 2012 to 2014, the number and types of DNA profiles being evaluated increased greatly from the days of manual interpretation. The mathematics was published in peer reviewed scientific journals and showed quite extensive testing and validation (many of which are given in this thesis) and direct attack by defence or defence experts on the evaluation of the DNA profile data became more and more difficult. This caused the nature of defence arguments to shift to two areas:

1) Conceding the presence of DNA, but disputing the mechanism of transfer which lead to its deposition on the item of interest (which is expanded on more in chapter 8)
2) The initial assessment by the analyst of the number of contributors to the DNA profile

Repeated criticisms in court of the choice of the number of contributors made by the analyst was the practical drive for the work that lead to the publication in this chapter.

Despite the work being carried out several years ago, the mathematics for treating the number of contributors as a nuisance parameter has not yet been introduced into active casework. The reason for this lack of forward movement has largely been due to the resistance of the STRmix™ developers to statistical methods that act as ‘black boxes’. Since the initial introduction of STRmix™ to the forensic community is has always been maintained that the conceptual functioning of the method is teachable, understandable by the forensic community and defendable by them when challenged. While there is no expectation that non-developing analysts using STRmix™ could derive all formulae in use, they know well enough the concepts to understand how the system works and importantly can identify when it has failed to work.

The mathematics behind the ability to choose a range of contributors is complex. The analysis requires the comparison of different posterior samples spaces of differing dimensions, and these must either be pitted against each other within an analysis (using

systems such as Reversible Jump MCMC), or compared afterwards. The workings of the calculation risks being black box if not properly implemented and taught. Much thought is required regarding the teachable and diagnosable elements of the calculation and various diagnostics that could indicate when an analysis has failed to work. This thought process is ongoing.

Manuscript: Interpreting forensic DNA profiling evidence without specifying the number of contributors. D Taylor, JA Bright, J Buckleton. (2014) Forensic Science International: Genetics 13, 269-280 – *Cited 5 times*

Statement of novelty: In this work, an MCMC posterior space comparison between analyses of differing dimensions, previously used in astronomy, was modified for use in DNA profiling problems.

My contribution: Main author, contributor to theory and sole simulation programmer.

Research Design / Data Collection / Writing and Editing = 70% / 100% / 60%

Additional comments: Also supplied is the supplementary material, that provides the derivation of the mathematics that allows the inter-dimensional parameter space comparison.

Forensic Science International: Genetics 13 (2014) 269–280

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

Original Research

Interpreting forensic DNA profiling evidence without specifying the number of contributors

Duncan Taylor [a,b,*], Jo-Anne Bright [c], John Buckleton [c]

[a] Forensic Science South Australia, 21 Divett Place, Adelaide, SA 5000, Australia
[b] School of Biological Sciences, Flinders University, GPO Box 2100, Adelaide, SA, 5001, Australia
[c] ESR, PB92021, Auckland, New Zealand

ARTICLE INFO

Article history:
Received 31 March 2014
Received in revised form 11 August 2014
Accepted 31 August 2014

Keywords:
DNA profile interpretation
Mixtures
Number of contributors
MCMC
Continuous model
STRmix

ABSTRACT

DNA profile interpretation has benefitted from recent improvements that use semi-continuous or fully continuous methods to interpret information within an electropherogram. These methods are likelihood ratio based and currently require that a number of contributors be assigned prior to analysis. Often there is ambiguity in the choice of number of contributors, and an analyst is left with the task of determining what they believe to be the most probable number. The choice can be particularly important when the difference between two possible contributor numbers means the difference between excluding a person of interest as being a possible contributor, and producing a statistic that favours their inclusion. Presenting both options in a court of law places the decision with the court. We demonstrate here an MCMC method of correctly weighting analyses of DNA profile data spanning a range of contributors. We explore the theoretical behaviour of such a weight and demonstrate these theories using practical examples. We also highlight the issues with omitting this weight term from the *LR* calculation when considering different numbers of contributors in the one calculation.

1. Introduction

DNA profile interpretation has benefitted from recent improvements that use semi-continuous (e.g. LRmix, LikeLTD, LabRetreiver) [1–5] or fully continuous (e.g. STRmix™, TrueAllele) [6–8] methods to interpret information within an electropherogram (EPG). These methods are likelihood ratio (*LR*) based and currently require that a number of contributors be assigned prior to analysis. Although it is possible in each of these systems to analyse the same profile under a number of different contributor options, the question still remains how to make use of the information. This is particularly true when the difference between two possible contributor numbers means the difference between excluding a person of interest (POI) as being a possible contributor, and producing a statistic that favours their inclusion. Presenting both options in a court of law places the decision with the court. If it is not possible for an expert to make this assignment it may be expecting a lot to ask the court to do so.

There are two recognised solutions to the problem, both of which have their foundation in the idea that the number of contributors is a nuisance variable and should be integrated out of the *LR* calculation. This is not a commonly held view. Many forensic biologists would consider the number of contributors to be something that should be determined from the data and hence are part of the output of the interpretation rather than a nuisance parameter that we will sum or integrate out. Budowle et al. state that "every effort should be made to provide the best estimate of the number of contributors" [9]. This, and many other statements of this type, gives voice to the commonly held, but incorrect, view that the number of contributors to a profile is knowable and is an important part of the output.

While other works have examined the issue of calculating *LR*s where the number of contributors is either unknown or can be bound in some sensible manner [10,11], the theory relies on having probabilities for the numerator and denominator of the *LR*, which simplifies the calculation. Commonly continuous systems, which utilise at least peak height information, work with probability densities. We explain here a Markov Chain Monte Carlo (MCMC) method for weighing different numbers of contributors against each other, and the practical consequences of including this information in the *LR*. In doing so we demonstrate that a single

* Corresponding author at: Forensic Science South Australia, 21 Divett Place, Adelaide, SA 5000, Australia. Tel.: +61 8 8226 7700; fax: +61 8 8226 7777.
E-mail address: Duncan.Taylor@sa.gov.au (D. Taylor).

http://dx.doi.org/10.1016/j.fsigen.2014.08.014
1872-4973/

exact number is not required, contrary to many currently held views.

1.1. Mathematics of the LR

The evidence obtained consists of a number of peaks (O^l_{ar} for allele '*a*' at locus '*l*' in replicate '*r*'), each with an associated height, molecular weight and DNA sequence. Whether this information is obtained from one or several replications of the same DNA sample the result is a series of observed peaks, which together are refer to as the observed data **O**.

We are interested in the probability of obtaining this observed evidence and in particular the probability of obtaining it given some competing propositions (or hypotheses), H_1 and H_2. We use Pr for probability and p for a probability density.

$$LR = \frac{Pr(\boldsymbol{O}|H_1)}{Pr(\boldsymbol{O}|H_2)}$$

In order to calculate these competing probabilities there are a number of nuisance parameters that we must consider, namely:

- The j genotype sets ($\mathbf{S}_j$) of contributors, each of which is made up of n single person genotypes for an n person mixture.
- The mass parameter (**M**), which is the term used for the grouping of parameters for template DNA amounts for each contributor (t_n), a degradation curve for each contributor (d_n), a replicate amplification efficiency for each replicated analysis (R_r), amplification efficiencies for each locus (A^l), and peak height variance constants for stutters and alleles.
- The number of contributors **N** to a DNA profile. Note that here **N** may signify one, or several possible numbers under consideration.

For many years the nuisance parameter that has been most concentrated is the genotype sets, which has been incorporated into the *LR* by:

$$\mathrm{LR} = \frac{\sum_j p(\boldsymbol{O}|\boldsymbol{S}_j, H_1) Pr(\boldsymbol{S}_j|H_1)}{\sum_{j'} p(\boldsymbol{O}|\boldsymbol{S}_{j'}, H_2) Pr(\boldsymbol{S}_{j'}|H_2)}$$

For many years simplifying assumptions have been made for $p(\boldsymbol{O}|\boldsymbol{S}_j, H_x)$, often to the point where these values where considered equal and simply removed from the equation, or designated as zero, resulting in the removal of the entire genotype set from the *LR* calculation. The likelihood, $p(\boldsymbol{O}|\boldsymbol{S}_j, H_x)$, is independent of the proposition because given a genotype set the likelihood will be the same regardless of whether a POI is being hypothesised as a contributor, i.e. $p(\boldsymbol{O}|\boldsymbol{S}_j, H_1) = p(\boldsymbol{O}|\boldsymbol{S}_j)$. In order to assess $p(\boldsymbol{O}|\boldsymbol{S}_j)$ it is helpful to introduce another set of nuisance parameters, the mass parameters. Following Taylor et al. [6] we denote these mass parameters **M**, and note $p(\boldsymbol{O}|\boldsymbol{S}_j) = \int_M p(\boldsymbol{O}|\boldsymbol{S}_j, \boldsymbol{M})\, p(\boldsymbol{M}) d\boldsymbol{M}$. We term $p(\mathbf{O}|\mathbf{S_j})$ as the 'weight' for genotype set j and give it the nomenclature w_j.

$$\mathrm{LR} = \frac{\sum_j w_j Pr(\boldsymbol{S_j}|H_1)}{\sum_{j'} w_{j'} Pr(\boldsymbol{S}_{j'}|H_2)} \quad (1)$$

Note the summations in the numerator and denominator of Eq. (1) may contain different numbers of non-zero elements, which we indicate by using summation indices of j and j'. We use this nomenclature throughout the remainder of the paper. To obtain weights we use the system of [6], whereby an MCMC process steps through different genotype sets from iteration to iteration. The weights obtained are residence times of each genotype set as the focus of the MCMC. In doing so we produce weights that sum to one, and are proportional to $p(\mathbf{O}|\mathbf{S_j})$. The value of $p(\mathbf{O}|\mathbf{S_j})$ is a density without its normalising constant, and we ask the reader to consider them as densities rather than probabilities as this assists between model comparisons discussed later in this paper. When considering the number of contributors, *N*, the weights must be considered as spanning models (consideration of different numbers of contributor). Note that Eq. (1) can still be thought of as having a term for number of contributors, however as the number is fixed for all components and there are no comparisons between models of different numbers of contributor then the *N* term is omitted. To obtain proportionality within and between models we note that the weights are products of a term for within model comparisons (which we will continue to use w_j) and a term for between model comparisons, which we term Z_n. Introducing Z_n into the *LR* gives:

$$\mathrm{LR} = \frac{\sum_n Z_n Pr(N_n|H_1) \sum_j w_j Pr(\boldsymbol{S_j}|N_n, H_1)}{\sum_{n'} Z_{n'} Pr(N_{n'}|H_2) \sum_{j'} w_{j'} Pr(\boldsymbol{S}_{j'}|N_{n'}, H_2)} \quad (2)$$

where N_n is a model specifying n contributors and we make explicit that there is no mathematical connection between the number of contributors under H_1, n, and under H_2, n'.

Details of the calculation and meaning of Z_n and the derivation of Eq. (2) are given in full in supplementary material 2 (note that the full derivation is mathematically dense, and the majority of this paper does not require an understanding to that depth).

This allows us to introduce the concept of dimensionality. There are more variables (dimensions) in the vector **M** for increased number of contributors. This means that the mass parameters are now dependent on contributor numbers so that $p(\mathbf{M}_n|N_n)$ is a probability density in multidimensional space. In MCMC parlance we could term the number of contributors different models and hence $N = 3$ could be one model and $N = 4$ another. There are a number of known ways to compare different models, some of which are within chain comparisons such as reversible jump MCMC [12]. The method used in this study is a between chain comparison by calculation of the marginal likelihoods.

1.2. The effect of propositions on the LR

1.2.1. The effect on genotype set weights

At this stage it is useful to discuss the propositions that are being considered. In Eq. (2) there are only two terms that depend on the propositions, $\Pr(\boldsymbol{S_j}|N_n, H_x)$ and $\Pr(N_n|H_x)$. The $\mathbf{S}_j$ term can be further decomposed into the genotypes of the known contributors (assumed to be present in the DNA sample by all parties, $\mathbf{S}_k$), the genotype sets of the POIs being postulated as contributors under one proposition but not all (typically in forensic contexts this will be a POI considered a contributor under H_1 but not in H_2, $\mathbf{S}_p$) and the genotypes of all the other, unknown contributors ($\mathbf{S}_U$) that must be present in the mixture to explain the total number of contributors to the profile that are not accounted for by known contributors or POIs. Using H_1 as an example, this gives:

$$Pr(\boldsymbol{S_j}|H_1) = Pr(\boldsymbol{S}_{U_j}|\boldsymbol{S_k}, \boldsymbol{S_p}, H_1) Pr(\boldsymbol{S_k}, \boldsymbol{S_p}|H_1) \text{ where } Pr(\boldsymbol{S_k}, \boldsymbol{S_p}|H_1) = 1$$

The proposition makes a difference here because out of all '*j*' possible combinations of n person genotypes that could describe **O**, many of them will not contain the genotype of the POI. In such a situation there are no combinations of genotypes of unknowns that can explain the data given $\mathbf{S_k}$ and $\mathbf{S_p}$, and so $\mathbf{S_U}$ is an empty set ($\boldsymbol{S}_{U_j} = \varnothing$) giving rise to $Pr(\boldsymbol{S}_{U_j}|\boldsymbol{S_k}, \boldsymbol{S_p}, H_1) = 0$. The remaining non-zero elements of $Pr(\boldsymbol{S}_{U_j}|\boldsymbol{S_k}, \boldsymbol{S_p}, H_1)$ are based on the rarity of the required genetic components (alleles) in whatever population is of interest.

Even though the same list of j genotype sets is considered under H_1 and H_2, with the same values of $p(\mathbf{O}|\mathbf{S_j})$, we often write the sum under H_1 as having j non-zero elements and under H_2 having j' non-zero elements. However because w_j are not dependent on

propositions, MCMC is able to be used to determine them based purely on the observed data, **O**.

1.2.2. The effect on contributor number weights

$\Pr(N_n|H_x)$ can be evaluated in one of two ways. If a proposition specifies exactly n contributors then:

$$Pr(N_{n*}|H_x) = \begin{cases} 1 & n* = n \\ 0 & \text{otherwise} \end{cases}$$

If there is ambiguity in the number of contributors that the propositions then a possible way forward is to consider all values for N_n equally likely:

$$Pr(N_{n*}|H_x) = \frac{1}{K} \quad \text{for all } n^*$$

where there are K different numbers of contributors being considered. However any values can be chosen for these probabilities if information exists to guide the choice.

1.2.3. The propositions themselves

Classically these propositions have always specified a number of contributors. However to treat uncertainty in the number of contributors, that number cannot be specified in the propositions. To allow uncertainty in the number of contributors it is necessary to consider propositions that do not specify an exact number of contributors, such as '*the POI is a contributor of DNA to this sample*' considered against '*the sample has originated from people unrelated to the POI*'.

Alternatively propositions can include contributor numbers as long as they encompass a range. In this instance the propositions become '*the POI and n to n' unrelated individuals are the sources of DNA*' considered against '*(n + 1) to (n' + 1) individuals, unrelated to the POI, are the sources of DNA*'. Both sets of hypotheses produce equivalent *LR*s.

1.3. Dealing with a range of contributors

We discuss the two options for dealing with a range of contributors within the *LR*. These are:

i. integrating the number of contributors out and
ii. assigning a number based on maximising $\text{pr}(\mathbf{O}|H_x)$ for each H_x.

which are explored in Sections 1.3.1 and 1.3.2 respectively.

1.3.1. Integrating out the number of contributors

This approach seeks to implement Eq. (2). We note that the total number of contributors to the profile and the genotypes of the unknown contributors to the profile are all nuisance variables. This is not a new concept and was discussed by Buckleton et al. [13] although given how obvious the equation is it is likely it was considered much earlier. There has been minimal uptake in the forensic community presumably due to a lack of software options that are able to implement it, and the difficulty in assigning the between model weights, Z_n.

Under this scenario the number of contributors is not specified under either propositions and rather a range is considered. Although informative prior probabilities for the number of contributors can be used it is more likely that an uninformative prior will be used so that the $\Pr(N_n|H_x)$ terms in Eq. (2) cancel each other out to give:

$$\text{LR} = \frac{\sum_n Z_n \sum_j w_j Pr(\boldsymbol{S_j}|N_n, H_1)}{\sum_{n'} Z_{n'} \sum_{j'} w_{j'} Pr(\boldsymbol{S_{j'}}|N_{n'}, H_2)}$$

1.3.2. Using the most probable number of contributors for each hypothesis (MPN)

This approach assigns the number of contributors (n) as that choice that produces the maximum posterior probability for the EPG, $\vee_n Z_n \sum_j w_j Pr(\boldsymbol{S_j}|N_n, H_1)$ and $\vee_{n'} Z_{n'} \sum_{j'} w_{j'} Pr(\boldsymbol{S_{j'}}|N_{n'}, H_2)$.

Using this approach the number of contributors for each competing hypotheses within the *LR* is assigned in such a way that it optimises the posterior probability of the EPG for that hypothesis. The number of contributors may be the same or different for each hypothesis, however a single number is chosen for each.

There are two competing forces that determine which choice of contributors is most favourable for a given EPG. Firstly there is a drive to minimise the number of unknown contributors under a hypothesis as each additional unknown contributor incurs an additional genotype probability in the calculation. This is true in the calculation of $\Pr(\boldsymbol{O}|H_x, \boldsymbol{M_n}, \boldsymbol{S_j}, N_n)$ and also in the calculation of Z_n. The number of imbalances or stochastic effects will drive an increase in the number of contributors as the profile will be described better with an additional contributor accounting for the imbalances.

The effect of each of these two competing components will dictate which choice of number of contributors produces the greatest posterior probability for the EPG. The *LR* calculated will ultimately be

$$\text{LR} = \frac{Z_n \sum_j w_j Pr(\boldsymbol{S_j}|N_n, H_1)}{Z_{n'} \sum_{j'} w_{j'} Pr(\boldsymbol{S_{j'}}|N_{n'}, H_2)}$$

where n and n' can be the same or different.

1.3.3. Choosing stratification or MPN

There are scenarios where a range may be more applicable than an MPN estimate for the choice in number of contributors, or vice versa. Note that the choice between stratification and MPN only affects the $\Pr(N_n|H_x)$ terms in Eq. (2) as outlined in Section 1.2.2.

Consider an intimate swab from a victim which has yielded a DNA profile with peaks at heights which could be reasonably explained by two contributors. The victim says she was raped by one man and has not had recent consensual sex. The main contributor to this profile corresponds with the victim, who is an assumed contributor. There are a number of minor alleles present all except one of which can be accounted for by the person of interest. Under the assumption of two contributors (and ignoring the possibility of drop-in for this scenario) the suspect would be excluded as a contributor of DNA to this profile. This is the position that defence may wish to take. The prosecution would take the stance that the profile has originated from three individuals and so would not exclude the suspect as a source of DNA. Under this scenario the MPN estimated values for number of contributors is arguably the better treatment of the profile.

Now consider a complex, low level DNA profile that has originated from an item where no-one can be assumed to have contributed DNA and there are several persons of interest for comparison. The profile can be described as two person profile, although there are indications that it may be from more than two, such as sub-threshold peaks, imbalances or drop-ins/drop-outs. Under this scenario, given the inability to reasonably assign a number of contributors to the profile, stratifying across a number of contributors is arguably the more appropriate choice.

1.4. The interaction of peak height variability and the number of contributors

There are four sources of ambiguity in assigning the number of contributors. These are:

1. Contributors sharing alleles, known as 'masking'
2. Artefactual peaks in allelic positions
3. Drop-out of alleles
4. Variability in peak heights

The masking of alleles has dominated considerations of the number of contributors [13–17], probably incorrectly. It is a common claim in court by defence experts the 'true' number of contributors to a profile could be different from the number used in statistical analyses. This of course misses the point that for evidence samples the 'true' number of contributors can never be known and is not required for *LR* calculations. It is also often forgotten that defence and prosecution have every right to nominate numbers of contributors in their own propositions, but have no jurisdiction over the other party's choice.

Our experience suggests that the interpretation of small peaks in forward $(a+1)$ stutter positions, larger than expected peaks in back stutter $(a-1)$ positions, and peaks imbalances are larger sources of ambiguity. The drop-out of alleles and variability in peaks heights are manifestations of the same underlying phenomenon, that the peak height of an allele is not directly related to the template available in the extract.

In current practice the assignment of a number of contributors usually proceeds by assigning peaks as allelic or artefactual. There may be ambiguity in the assignment of peaks as artefactual and this arises most often when backstutter peaks are of a similar height to some unambiguously allelic peaks. This uncertainty is mentally "carried forward." A putative minimum number of contributors is then typically assigned as the maximum number of allelic peaks divided by 2. If this is not a whole number the result is rounded up. This putative number is then subjectively trialled against the EPG for potential fit to peaks heights and the ambiguously artefactual peaks.

Variation in peak heights has historically been treated using a threshold of acceptance or rejection for the ratio of two peaks from a heterozygote (*Hb*). Thresholds on *Hb* may be soft or hard. Consider, for example, two peaks of height 2000 and 1000RFU. The ratio of these is 2:1 and this ratio may be considered high if the sample is from a single donor. This would suggest that the inclusion of an additional contributor is warranted. This mental process is informally considering the likelihood of the observed data given a number of contributors. What is commonly lacking in this consideration is the effect of the genotype probabilities when adding a contributor. The prior probability of the genotype sets drops dramatically in some circumstances due to the multiplication of an additional profile frequency in the prior. In Example 1 the difference between likelihood and posterior probability (due to changes in prior probabilities) is demonstrated.

Empirical studies suggest that the variance in peak height is inversely proportional to the amount of template [18–21]. We model the system's tolerance to stochastic effects using a variance constant term. See [6] for a full explanation of the model, but to surmise the observed peak height (O) is compared to the expected peak height (E) using the model:

$$\log\left(\frac{O}{E}\right) \sim N\left(0, \frac{c^2}{E}\right)$$

where c^2 is the variance constant. This variance constant is not known with certainty and may vary between different samples and between models. We model the constant as having a gamma distributed prior. This choice appears reasonable from empirical data. However, following Balding [3] we allow the profile under consideration to influence the value for this constant.

Consider an apparently single source sample based on allele count. When treated as a single source profile the difference in peak heights within and between loci has to be treated as variance. If the variance is low (an intolerant variance) then a low probability will be produced for any imbalanced peaks. If the variance is high (a tolerant variance) then the system will produce moderate probabilities for imbalanced peaks.

We now consider the effect of adding a second contributor to a potentially single source profile. Certain combinations for this second contributor can improve the fit of the proposed genotypes to the profile. This will have the biggest effect for an intolerant variance and a lesser effect for a tolerant variance. We would therefore expect that the addition of a second contributor would improve the likelihood of the profile most when using an intolerant variance, or when large imbalance exists within the profile. This is also demonstrated in Example 1.

1.5. The effect of different EPGs will have on Z_n

1.5.1. If peak heights are high and balanced

The consideration of a number of contributors above the minimum required to reasonably describe the profile will not substantially improve the fit to the profile. In this instance the additional unknown contributor is likely to be considered as a very trace contributor in a probabilistic analysis.

This is contrary to many expectations that adding contributors to only one proposition's case is always detrimental to that case. For a mathematical demonstration of this concept we point the reader to Supplementary material 1. We provide a practical demonstration in Example 2. Note that this almost independence of the *LR* from the number of contributors is only true under the specific circumstances that the POI can account for one of the dominant contributors to the profile and that the addition of contributors provides no improvement to the description of the observed profile. When one or both of these is not the case the effects of changing numbers of contributors can be dramatic, as shown in Example 3 and Section 4.2.

1.5.2. If the peak heights are high and imbalanced

A choice of number of contributors that can account for the imbalances in some reasonable manner will describe the observed EPG much better than fewer contributors. In this instance the two effects of genotype probability and profile fit will be acting against one another and the weight will depend largely on the severity of the imbalance and rarity of the alleles. Example 3 explores this concept.

1.5.3. If peak heights are low

The scenario will be similar to the balanced profile considered above. The *LR* generated will depend on the balances within the profile, the intensity of the peaks, and rarity of alleles.

Example 1. The effect of the model on Z_n

Before we can assess a system's ability to generate weights that correspond to numbers of contributors we must ascertain how that system should be set up. Consider a strong, balanced single source profile (Fig. 1). We will interpret this as single source (correct) and as a two person mixture (incorrect).

Changing the shape of the variance constant prior can make the method more or less tolerant of stochastic effects. This has a direct effect on the values of Z_n, as imbalances that were acceptable given a more tolerant system, become vastly less probable.

The gamma prior distribution for the variance has two variables, shape and scale. Gamma distributions were chosen to produce a narrow distribution about a desired mode. Modes tested

were 0.1, 0.5 and 5. The narrowness of the distribution means the variance constant value will be very limited in how much the profile in question can affect it. This allows the effect of the variance and the number of contributors to be viewed with less confounding of effects.

For all calculations of *LR*s or Z_n values in this paper, allele frequencies were used from an Australian, self-declared Caucasian database [22].

Fig. 2 demonstrates that reducing the variance constant to very low levels leads to a marked favouring of the addition of a contributor to describe the profile. This is because at these levels the system is highly intolerant of imbalances, and so the slight differences between observed and expected peak heights are explained by the presence of a trace second contributor. Note that in Fig. 2 the distributions are only the likelihood of the observed data given the mass parameters within a model, they have not be multiplied by priors that include the allele frequencies. This can be seen by the distance between the modes of the distributions compared to the ratio of Z_n values being different by approximately the frequency of a full profile ($\sim 10^{12}$) when the variance constant is less than 0.5. This difference between distributions Z_n values is not seen when the variance constant is 5. The reason for the different behaviours can be explained by the ambiguity in the genotypes of the second contributor and the role that has to play in the calculation of Z_n.

Consider a profile originating from two individuals with a locus that specifies a mixture proportion and a variance constant that 'locks' that mixture proportion across the profile. For

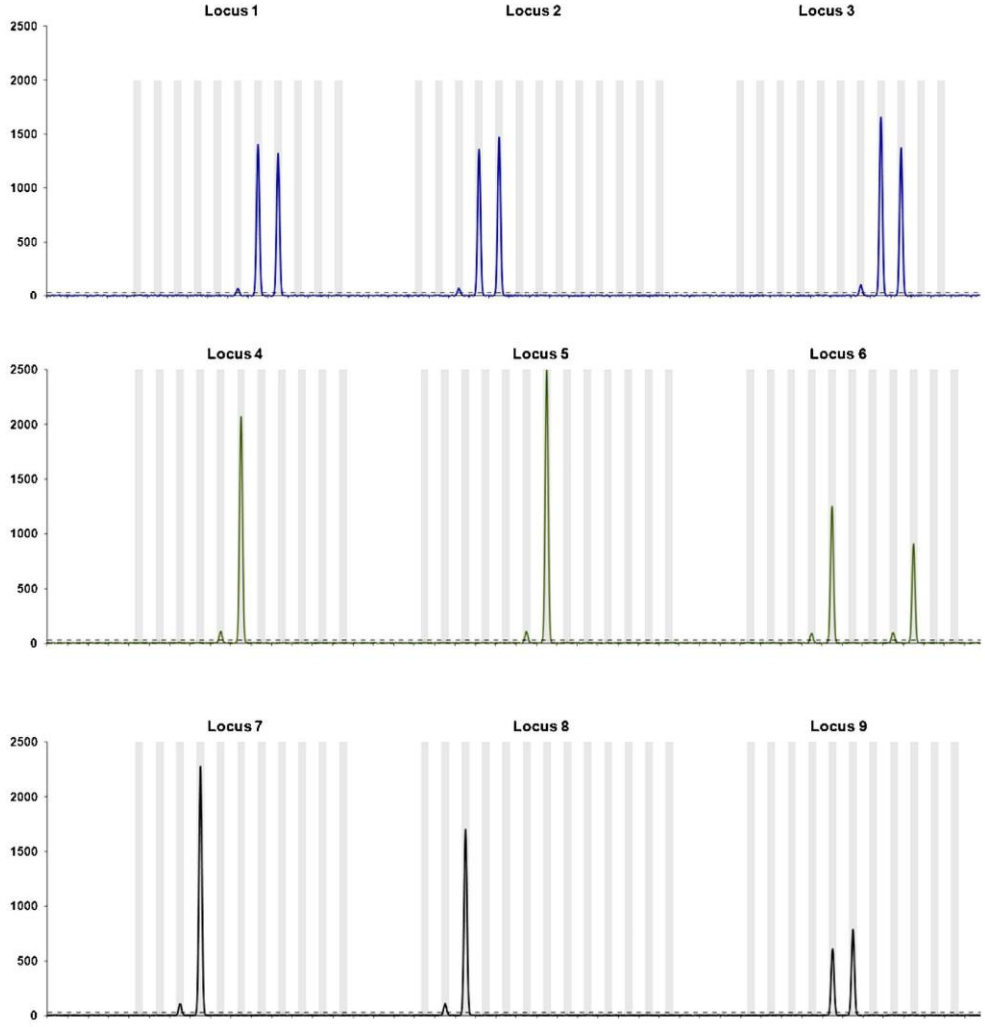


Fig. 1. Single source profile used for calculation.

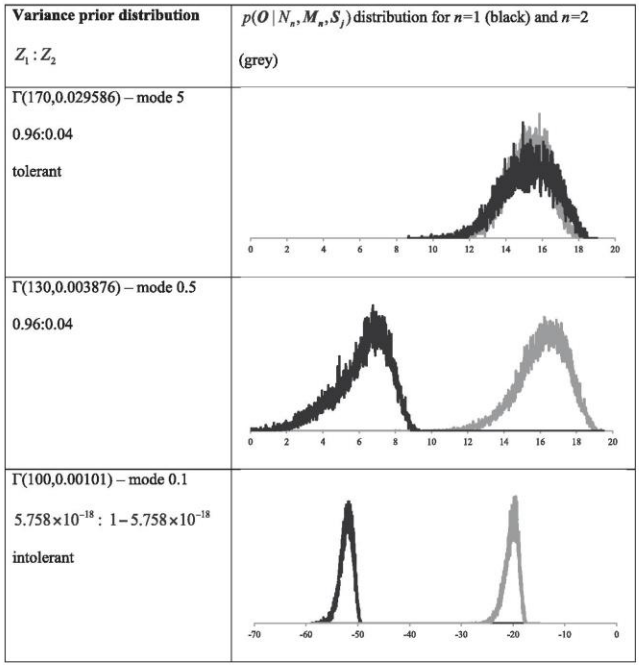


Fig. 2. Graphs showing distribution of $\log_{10}[\, p(O|N_n, \boldsymbol{M_n}, \boldsymbol{S_j})]$ on x-axis for profile seen in Fig. 1, when considered as originating from one (black) or two (grey) individuals. Note that these distributions are log likelihoods only and do not include priors i.e. for mass parameters, number of contrbutors or genotype set probabilities. Also note that these log likelihoods are produced by an MCMC process which steps through different genotype sets. The distributions produced have resulted from numerous genotype sets being the focus of the MCMC at various iterations during its run.

example picture a locus with two peaks [A,B] where [A] is 1000rfu and [B] is 250rfu. Imagine that the variance constant is small enough that only two genotypic explanations of the data are acceptable under N_2, and that is the genotypes are [A,A]&[A,B] with a mixture ratio of 0.6:0.4 or [A,A]&[B,B] with a mixture ratio of 0.8:0.2. If the profile was considered as originating from a single individual then the pairing of [A] and [B] will incur some penalty as the peaks are imbalanced i.e. $p(\boldsymbol{O}|N_2, \boldsymbol{M_2}, \boldsymbol{S_j}) > p(\boldsymbol{O}|N_1, \boldsymbol{M_1}, \boldsymbol{S_j})$. This will be offset by the difference in $\Pr(\boldsymbol{S_j}|N_n)$ within $\int_{M_n} \sum_j p(\boldsymbol{O}|N_n, \boldsymbol{M_n}, \boldsymbol{S_j})\, p(\boldsymbol{M_n}|N_n)\, p(\boldsymbol{S_j}|N_n)$ used to generate Z_n. Now consider that this is the only indication of a second contributor in the profile and we move on to the next locus, which is homozygous and has single peak [C] at 2000rfu. At this locus $p(\boldsymbol{O}|N_1, \boldsymbol{M_1}, \boldsymbol{S_j}) \approx p(\boldsymbol{O}|N_2, M_2, \boldsymbol{S_j})$ as the additional contributor does not increase the likelihood of the observed profile. The genotypes of the potential contributor under N_1 is [C,C] and under N_2 is [C,C]&[C,C]. There are no other acceptable genotype sets under N_2 as the first locus has dictated that the minor contributor is providing approximately 400 to 800rfu to [C] (based on the mixture proportion of 0.4–0.2). At this level dropout is sufficiently unlikely that it can be ignored. So in the calculation of Z_n $p(\boldsymbol{M_1}|N_1)p([C,C]) \gg p(M_2|\boldsymbol{N_2})p([C,C]\,\&\,[C,C])$ by approximately the frequency of a [C,C] genotype and the difference between $p(\boldsymbol{M_1}|N_1)$ and $p(\boldsymbol{M_2}|N_2)$ which is discussed later. This trend will continue across all loci where there is no improvement in fit from the addition of a contributor and the genotype sets are restricted by the mixture proportions.

If the penalties incurred from imbalances outweigh the rarity of an additional contributor's profile then $Z_2 > Z_1$ as seen in Fig. 2 at mode 0.1. As variance increases the imbalances are tolerated more and the penalty is lower, however if the variance is still small enough to restrict the genotype set then a point is reached where $\frac{p(\boldsymbol{O}|N_2, \boldsymbol{M_2}, \boldsymbol{S_j})}{p(\boldsymbol{O}|N_1, \boldsymbol{M_1}, \boldsymbol{S_j})} < \frac{Pr(\boldsymbol{S_j}|N_n)}{Pr(\boldsymbol{S_j}|N_n)}$ and this will result in $Z_2 < Z_1$ as seen in Fig. 2 at mode 0.5.

Now consider the same situation, but this time at locus 1 the peak height of peak [B] is 1010rfu (indicating some small level of imbalance that could be explained by standard stochastic effects). Imagine that at all loci the likelihood of the profile is not substantially increased by the addition of a second contributor, i.e. $p(\boldsymbol{O}|N_1, \boldsymbol{M_1}, \boldsymbol{S_j}) \approx p(\boldsymbol{O}|N_2, \boldsymbol{M_2}, \boldsymbol{S_j})$ at all loci. Additionally the second contributor is deemed to be providing very little template to the observed profile. Using the second locus again, if we consider the priors, and particularly the genotype probability prior then $\sum_j p(\boldsymbol{M_1}|N_1)Pr(\boldsymbol{S_j}|N_1) = p(\boldsymbol{M_1}|N_1)Pr([C,C])$, as there is still only one genotype that can explain the observed profile. In the two person scenario $\sum_j p(\boldsymbol{M_2}|N_2)Pr(\boldsymbol{S_j}|N_2)$ is calculated by:

$$= p(\boldsymbol{M_2}|N_2)\{Pr([C,C]\&[C,C]) + Pr([C,C]\&[C,Q]) + Pr([C,C]\&[Q,Q])\}$$
$$\approx p(\boldsymbol{M_2}|N_2)Pr([C,C])$$

And hence the difference between Z_1 and Z_2 will be based on the difference between $p(\boldsymbol{M_1}|N_1)$ and $p(\boldsymbol{M_2}|N_2)$. Under these circumstances Z_1 will be mildly favoured over Z_2 as $p(\boldsymbol{M_2}|N_2)$ contains an

extra template term and an extra degradation term (or dimension), each with a prior. This result can be seen in Fig. 2 mode 5 where Z_2 is mildly less than Z_1, and demonstrates the theory that MCMC systems favour simplistic models.

For the remainder of the paper a variance constant prior gamma distribution of $\Gamma(1.62, 3.98)$ was used for alleles and $\Gamma(2.57, 3.57)$ for stutter peaks. It is known that stutter and allele peaks have different peak variance values [20]. These values were optimised to Profiler Plus control data (analysis not shown).

Example 2. Considering different dimensions in H_1 and H_2

We consider the results from Example 1 but this time consider it as originating from one, two, three or four individuals. The known source was used as the POI and was then compared with the analysis at each stage using propositions:

H_1: POI + $(n-1)$ unknowns

H_2: n unknowns

where n can be one, two or three and n' can be one, two, three or four and n is independent of n'. Table 1 shows $\sum_j w_j Pr(\boldsymbol{S_j}|N_n, H_1)$, $\sum_j w_j Pr(\boldsymbol{S_j}|N_n, H_2)$ and Z_n for the four contributor scenarios.

Note a mild decrease in likelihoods as the number of contributors increases, even without the Z_n terms. This demonstrates the ability of the continuous MCMC method to overcome one of the problems associated with purely probabilistic systems that work by maximum likelihood estimation, which is that an increase in the number of contributors always increases likelihood (see Section 5.1 of [23], who demonstrates this phenomenon). Table 2 shows the $\log_{10}$(LR) considering different combinations of individuals in H_1 and H_2 when it is assumed that $Z_n = Z_{n+1}$, i.e Z_n from Table 1 is not included in the *LR* calculation.

Table 3 is a repeat of Table 2, but this time including Z_n in the *LR* calculation.

Table 1
Probabilities densities of the observed profile in Fig. 1 given a varying number of contributors and the Z_n values associated with those numbers of contributors.

n	$\sum_j w_j Pr(\boldsymbol{S_j}\|N_n, H_1)$	$\sum_j w_j Pr(\boldsymbol{S_j}\|N_n, H_2)$	Z_n
1	6.76×10^{23}	1.87×10^{13}	0.8087
2	1.74×10^{22}	4.8×10^{11}	0.0818
3	3.23×10^{20}	8.23×10^{9}	0.0227
4	7.28×10^{19}	2.01×10^{9}	0.0868

Table 2
$\log_{10}$(LR) considering differing number of contributors under H_1 and H_2.

		Contributors under H_2			
		1	**2**	**3**	**4**
Contributors under H_1	**1**	10.6	11.6	12.6	12.8
	2	9.2	10.3	11.3	11.5
	3	8.0	9.1	10.1	10.3
	4	7.7	8.8	9.8	10.0

Table 3
$\log_{10}$(LR) considering differing number of contributors under H_1 and H_2 and including Z_n.

		Contributors under H_2			
		1	**2**	**3**	**4**
Contributors under H_1	**1**	10.6	12.6	14.2	13.8
	2	8.2	10.3	11.8	11.4
	3	6.5	8.5	10.1	9.7
	4	6.7	8.8	10.4	10.0

In this instance the change in *LR*s caused by including Z_n is slight. This is because the addition of contributors in the model beyond one, does not significantly improve the description of the observed peak heights. Additionally there is ambiguity in the genotypes of the additional (unnecessary) contributors, such that there are many possibilities, including complete dropout, that they can take. The sum across these genotypic probabilities is high and therefore has a small impact on Z_n. The results seen in Tables 2 and 3 demonstrate the theory shown in Section 1.5 that under the tested circumstances the *LR* should remain reasonably constant regardless of the addition of contributors under both or either one of the propositions. It should also be noted that due to the slight favouring of simpler (lower contributor) models, there is still no advantage in artificially increasing the number of contributors to one or both of the hypotheses as this will tend to drive the *LR*'s support away from the proposition with the greater number of contributors.

If the number of contributors increases under both propositions (moving down the diagonal of Tables 2 and 3), the *LR* decreases, but not due to any loss of resolution in the 'major' contributor's genotype. Note that the *LR* for the two, three and four person scenarios are approximately one half, one third and one quarter the *LR* of the single source equivalent (note the $\log_{10}(LR)$ if you are comparing $10^{10.6}$, $10^{10.3}$, $10^{10.1}$ and 10^{10}). This is because the propositions being considered do not nominate a specific contributor position for the POI (see [24] for a full explanation of the concept).

Example 3. Imbalanced peaks

We use again the profile given in Fig. 1, but introduce imbalance at locus 1. The height of the second allele at locus 1 is artificially adjusted in the range 40 to 1322rfu (its original height). The peak heights assigned to peak 2 in locus 1 were 40, 250, 500, 750, 1000 and 1322rfu, which produced heterozygote balance *Hb* values that fill the range between zero and one. The profile was again analysed as either originating from one or two contributors. Fig. 3 shows the ratios of the Z_2/Z_1 across the range of *Hb* and demonstrates the effects of the relationship between $p(\boldsymbol{O}|N_n, \boldsymbol{M_n}, \boldsymbol{S_j})p(\boldsymbol{M_n}|N_n)$ and $\Pr(\boldsymbol{S_j}|N_n)$ within $\sum_j p(\boldsymbol{O}|N_n, \boldsymbol{M_n}, \boldsymbol{S_j})\, p(\boldsymbol{M_n}|N_n) Pr(\boldsymbol{S_j}|N_n)$ as described in Example 1. At this point we omit the possibility of drop-in from the calculation as we are interested in showing only the effects of the imbalance on the values of Z_n:

There are a number of observations that can be made from Fig. 3:

- $Hb < 0.2$—the severity of the imbalance means that two contributors explains the profile sufficiently better to outweigh the additional profile probability. At the most extreme *Hb* the height of the less intense peak means that at other loci, when considered as a two person mixture, the genotypes the second contributor can take include complete dropout. This further adds support to the two contributor model.
- $0.2 < Hb < 0.4$—as *Hb* increase the imbalance penalty decreases, and the genotype sets for the second contributor are constricted by the mixture proportion, determined by the imbalance at locus 1. The additional contributor's genotype frequency outweighs the penalty from the imbalance at locus 1. This relationship reaches the pinnacle at $Hb = 0.35$.
- $0.4 < Hb < 0.8$—the penalty from the imbalance is still steadily decreasing, however the difference between the two peaks means that the optimal contribution of the additional contributor is becoming less. Consequently, moving along the *x*-axis from *Hb* of 3.5–5.5 the genotype set weights are being spread over more possible genotypes.

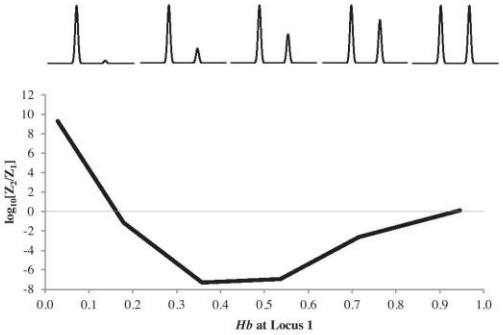


Fig. 3. Improvement in model description obtained by an additional contributor as a function of Hb, with Hb shown diagrammatically above.

- $0.8 < Hb$—after this point the penalty from Hb is relatively minor. The mixture proportion of the second contributor is optimised to close to 0% and so there is complete ambiguity in their genotype. The result is that $Z_1 \approx Z_2$, with Z_1 only being slightly higher due to the presence of an additional template and degradation prior under N_2.

We can consider how this might affect an *LR* calculated where the comparison is to a POI who is homozygous at the first locus i.e. the presence of the second peak at locus 1 needs to be described as either artefactual (e.g. drop-in, which we are not considering) or a second contributor for the POI to produce an $LR > 0$. The classical treatment of this problem would be for the two competing scenarios to nominate their number of contributors and then for the *LR* to be calculated giving equal weight to the numbers of contributors. Providing equal weight to either scenario could substantially disadvantage the case where an additional contributor provides a better, or worse, explanation for the observed profile. We show the effect of this in Fig. 4 by comparing the *LR*s produced by assuming $Z_1 = Z_2$ (solid line) i.e. Z_n effectively terms are not included in the *LR* calculation, and by using the informed Z_n values (dashed line), from the results seen in Fig. 3.

We also include in Fig. 3, a comparison to systems that do not take peak height into account. Under such a system each genotype set weight is composed of fixed probabilities of dropout (which we set to 0.05) and drop-in (which, to be consistent with other systems being displayed, we set to 0). As peak height is not taken into account then by definition changing the height of one of the peaks should have no effect on the *LR*, hence the horizontal line seen in Fig. 4.

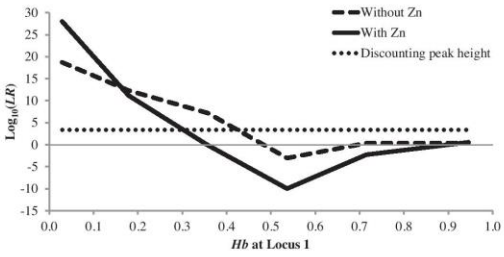


Fig. 4. LR with and without Z_n and also discounting peak height information.

- When $Hb < 0.2$, the profile strongly supports N_2 and if this is not taken into account the prosecution case is unfairly disadvantaged. Note that the *LR* is actually greater than the inverse of the profile frequency as in this example we have forced H_2 to consider only one contributor, which is a very poor description of the observed profile, and this low genotype set weight is used in the *LR* under H_2.
- In the range $0.2 < Hb < 0.8$ the additional contributor model is not supported by the data and omitting Z_n terms disadvantages the defence. A point worth noting is $0.4 < Hb < 0.5$ where the classic treatment favours inclusion of the POI slowly trending down to $LR = 1$, (note that by random chance variation the MCMC analysis has resulted in the dotted line dipping below the line of equality at $Hb = 0.54$, but the general trend expected would be for the *LR* to slowly to decrease to one as Hb increases to one) and including the Z_n values results in an *LR* favouring exclusion of the POI, in some cases by many orders of magnitude.
- When $Hb > 0.8$, then the profile can again be reasonably considered single source, but this time not matching the POI (at locus one). The additional contributor in H_1 adds no further fit to the observed profile. As $Z_1 \approx Z_2$ and there is complete redundancy in the additional contributor's genotype then the rarity of the single unknown's genotype in H_1 approximately equals the rarity of the single unknown's genotype in H_2 and the *LR* is driven towards one.

The difference in *LR* between the dotted line and solid line in Fig. 4 therefore represents the information being lost when assuming $Z_1 = Z_2$ in the *LR* calculation. Using informative weightings for number of contributors rather than uninformative can make a significant difference to the *LR*.

Most existing treatments of trans-contributor problems use uninformative weights for contributor number, and so are not appropriately assessing the two scenarios against each other. The alternative is to require a human interpretation and some system of conventional thresholds to ensure that the choice of number of contributors under both propositions can reasonably describe the observed EPG(s).

2. Validation of model

2.1. Variability in Z_n

The values for Z_n are determined using the system described from a MCMC system and so are subject to run to run variation. We investigate the variability in Z_n by running each of the analyses used the generation of Fig. 3 five times. The average number of post-burn-in iterations was 5.94×10^5 when considered a single source profile and 2.18×10^6 when considered a two person mixture. We carried out the same analyses as lead to Fig. 3 again in Fig. 5 but with all values for $\log_{10}(Z_2/Z_1)$ displayed, and the trend line showing the average $\log_{10}(Z_2/Z_1)$ value.

Fig. 5 shows that running the MCMC in the way described produced estimates for Z_n values that span approximately three orders of magnitude at the most variable Hb values. Note that the remainder of Section 3.1 delves into the components that make up Z_n in order to identify the source of the variation seen in Fig. 5, and hence an understanding of the contents of Supplementary material 2 is required.

Let the maximum likelihood value of posterior sample of the MCMC $p(\boldsymbol{O}|N_n, \boldsymbol{M_n}, \boldsymbol{S_j})\, p(\boldsymbol{M_n}|N_n) = \vee_{M_n,n}\, p(\boldsymbol{O})$, which selects the set of mass parameters that best describes the observed profile under each number of contributors. Ideally this value would be close to the maximum possible theoretical mode of the problem, however as the mode can be very 'pointy' [25] it is not always reached. Therefore the value of $\vee_{\boldsymbol{M_n},n}\, p(\boldsymbol{O})$ varies from run to run by approximately one to three orders of magnitude. This value is

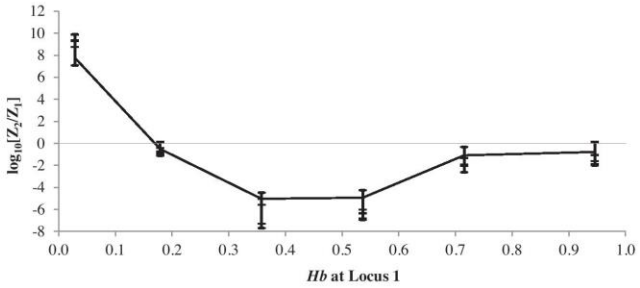


Fig. 5. Fig. 3 reproduced 10 times to show reproducibility. The line intercepts the average value for each Hb bracket.

also directly proportional to the final Z_n values as seen in (data not shown). These results suggest that long runtimes may be necessary in some instance for MCMC to get into the hyper-dimensional sample space close to the theoretical maximum mode. Also, due to the additional dimensions when considering a higher number of contributors, reaching the theoretical mode is likely to take more iterations as the model becomes more complex.

This suggests that much extended runtimes may still be necessary in some instances despite our best efforts. This significantly affects the practicality of the method if it is to be used in a forensic casework setting where significant time pressures commonly exist.

2.2. Test on complex data

The experiments shown in Example 3 indicate that a single extreme imbalance, or a number of mild imbalances, can drive the support for a higher number of contributors ($Z_n < Z_{n+1}$). However in Example 3 the imbalance needed to be very large in order for a higher number of contributors to be strongly favoured and in reality this level of imbalance would give cause to a scientist to choose that higher number unambiguously anyway via human interpretation. We have therefore demonstrated the theory but shown limited practical advantage. We turn now to a more complex scenario. Fig. 6 shows a complex profile originating from three individuals in proportions 0.17:0.42:0.42. This profile has been constructed so that peak masking and dropout means that by peak counting alone the mixture could be described by two contributors with some imbalances, high stutter and mixture proportion inconsistencies across the profile.

The MCMC analysis was run for 2×10^6 iterations under N_2 and 5.4×10^6 iterations under N_3. The likelihood ratios calculated by comparison to the known contributors are given in Table 4 when using propositions:

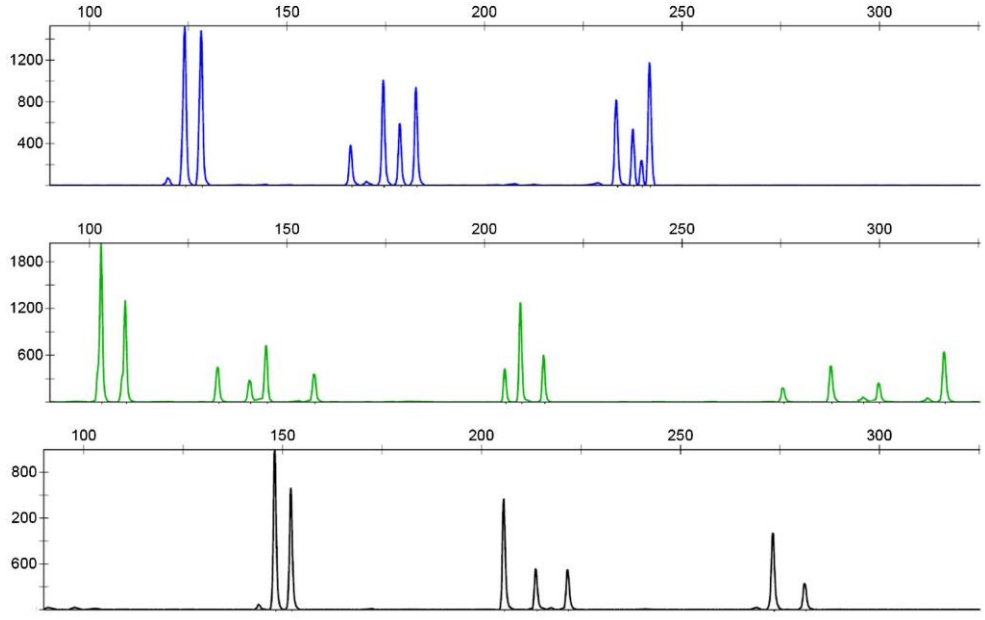


Fig. 6. Complex three person mixture masquerading as a two person mixture.

H_1: POI + $(n-1)$ unknowns

H_2: n unknowns

where n was two or three.

The values for Z_n are $Z_3 = 0.94$ and $Z_2 = 0.06$ indicating support for the three person scenario. To some analysts this may not be as much support for N_3 as expected, there are two reasons for this. Firstly the addition of the third contributor improves the fit to the observed data and consequently they do not have complete redundancy in the genotypes which they can take. This restricted genotype set prior probability therefore has an effect on Z_3:Z_2. Secondly we allow the peak variance constant to be optimised by the profile under N_2 model. Therefore the poor fit to the observed data has increased the variance constant and consequently become more tolerant of imbalances. In effect this analysis is telling us that the profile could be a poor fitting two person mixture with a high peak height variability or a better fitting three person profile with a lower peak height variability. Again the Z_n values appropriately weigh these scenarios against each other in the *LR* calculation.

We now calculate the *LR*s for known contributors using MPN for number of contributors:

H_1: The POI and $(n-1)$ unknowns are the sources of DNA

H_2: n' unknowns are the sources of DNA

(where n can be either 1 or 2 and n' can be either 2 or 3)and stratification across the range of contributors (2–3):

H_1: The POI and 1 or 2 unknowns are the sources of DNA

H_2: 2 or 3 unknowns are the sources of DNAand display the results in Table 5.

Profiles that masquerade as lower order mixtures can strongly favour their correct number of contributors as there will be multiple loci with imbalances or deviations from consistent mixture proportions across a profile. These combined effects can strongly drive Z_n towards the higher order scenario as seen in this example.

The profile seen in Fig. 6 may not be immediately recognised as a three person mixture, particularly if the DNA profiling system is known to have high peak height variability. However, incorrectly treating this profile as two person mixture excludes the known contributors as seen in Table 4. When analysed under the correct number of contributor all known contributors give much larger *LR*s and none are excluded, as would be expected. However the scientist cannot (and certainly should never) base their decision on number of contributors purely so that comparison of the POIs in a case yield an $LR > 1$. The analyst is left with a choice as to whether the stochastic events within the profile provide enough evidence to analyse it as a three person mixture, potentially overestimate the number of contributors required and falsely include a non-contributor. Or the analyst could surmise that the stochastic events do not warrant a third contributor, and so analyse it as a two person mixture and potentially falsely lower the *LR* or exclude true contributors. In some forensic laboratories, current practise would be for the analyst to analyse the profile as originating from three people under H_1 and two under H_2. The calculation of the *LR* would then be carried out without any weight terms for the number of contributors. Alternatively some laboratories would analyse the profile as a two person mixture under both propositions and, then analyse the profile as a three person mixture under both propositions and report both *LR*s. Neither of these two options appropriately weights the evidence. It is this instance that the Z_n values appropriately deal with the ambiguity in number of contributors.

It can be seen in Table 5 that either the MPN or stratification methods for dealing with ambiguous number of contributors gives an appropriate statistical weighting to the evidence for each known contributor, and is the same as the values given in Table 4 at the significance level show. The combination of Z_3/Z_2 and w_j means that N_3 is chosen using the MPN method for both H_1 and H_2, hence the values being the same as in Table 4. In the stratification method $p(\boldsymbol{O}|N_3, \boldsymbol{M}_3, \boldsymbol{S}_j)$ dominates the *LR* and so again at the significance level show the values in Table 5 are the same as in Table 4.

Table 4
Results from analysis of constructed three person mixture when analysed as $n=2$ and $n=3$.

	LR under N_2	LR under N_3
POI 1	0	1.03×10^6
POI 2	0	3.0×10^5
POI 3	0	2.6×10^8

Table 5
Results from analysis of constructed three person mixture when considered as a two to three person mixture.

	MPN	Stratified
POI 1	1.03×10^6	1.03×10^6
POI 2	3.0×10^5	3.0×10^5
POI 3	2.6×10^8	2.6×10^8

3. Conclusion

The advent of more sophisticated techniques of analysing DNA profiles has led to informative statistical weighting being obtained from a greater number of DNA profiles. Currently most systems of DNA profile interpretation require that a number of contributors be set prior to analysis. To do this the analyst must rely on their knowledge of DNA profile behaviour and make assessments on whether certain stochastic events are acceptable given a posited number or contributors. The acceptability of stochastic events requires rules and thresholds (even if values are not specified exactly), which is thorn in the side of advocates of continuous DNA interpretation systems that are designed to remove all such thresholds.

We attempt here to extend the model of [6] so that a posterior probability of a number of contributors to a model can be obtained as part of the MCMC process being used to analyse the profile. Obtaining these relative weights for profile dimensionality, Z_n, allows the calculation of an *LR* where either different numbers of contributors are posited under the two propositions, or a range of contributors is desired using weights informed by the data itself. This is an advance to the most common current method of dealing with trans-contributor analyses, which is either to:

(1) increase the number of contributors under both propositions or
(2) calculate the *LR* with a different numbers of contributors under the different propositions, but not taking into account the relative fits of these models to the observed profiles.

We have shown that both of these options can bias the result by many orders of magnitude. The first is biased when the increase in number of contributors is not required under one of the propositions (typically H_2) and the second can be biased when a number of unlikely stochastic events are required to explain the observed EPG(s) under a specific value of n.

Conceptually this work provides examples that show DNA profile evidence can be analysed assuming different numbers of contributors, and the results of these separate analyses can be combined by stratification for each proposition, using a model

weight, Z_n. The generation of Z_n will depend on how well each model describes the observed data, but also encompasses the additional profile frequencies required by the addition of contributors. The balance of these two factors ultimately dictates when one model is favoured over the other.

DNA profile analyses present some difficult challenges for MCMC methods, which other applications do not, namely:

- Time pressures for active casework results, typically results are desired with only minutes of runtime.
- Consistency of results between runs, especially challenging given the point above as the typical solution to this is to run the MCMC analysis for more iterations.
- A huge number of values that categorical parameters (genotype sets per locus) can take.
- The process must be able to work on a range of profile complexities using standard settings that do not require tuning from profile to profile.
- The process cannot be a 'black-box' as biology analysts with a wide variety of training (which usually does not include mathematics or statistics) must be able to implement the mathematics and explain the results in court.

Given these pressures we trial here calculating the Z_n values by calculation of the relative likelihood of one model over the other given the data using the method of Weinberg [25,26] which performs in an intuitively sensible manner given the tests we have run. Given all criteria mentioned above the method has worked well but we recognise the following issues:

(1) The process can require a number of MCMC iterations that may be impractical in forensic laboratories if using a range of contributors were to become a common event, rather than an exception. There is evidence of substantial variation in the results displayed in Fig. 5, although further investigation is already underway is aimed at reducing this variability further without requiring longer MCMC runs.
(2) The mathematics is approaching the complexity of a black-box. However, the conceptual introduction of a weight for the number of contributors is intuitively sensible and mathematics may not need to be fully understood to present these results in court. We have tried to structure the paper with this in mind, allowing readers to digest the majority of the paper without the need to refer to the complex mathematics in Supplementary material 2. It is thought by the authors that a conceptual level of understanding of what Z_n represents would be sufficient for almost any court challenge.

Another issue, although not relevant directly to the mathematics described in this paper, that will need to be addressed is one of communicating DNA results to court. Courts in the authors' countries have been used to scientists providing opinions on a number of contributors to a profile. Using a range of contributors indicates something which DNA statisticians have realised for some time but not had the means to act on, which is that the number of contributors to a DNA profile is not known and can never be known for most forensic casework. Further, the number of contributors does not need to be known to assess DNA evidence as long as the ambiguity in this number can be accounted for appropriately in the statistical model.

Acknowledgements

This work was supported in part by grant 2011-DN-BX-K541 from the US National Institute of Justice. Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the U.S. Department of Justice. We would like to thank Ian Evett (Principal Forensic Services), Martin Weinberg (Department of Astronomy, University of Massachusetts) and Darfiana Nur (School of Computer Science, Engineering and Mathematics, Flinders University) for helpful discussions and contributions to this work. Finally we would like to thank two anonymous reviewers, whose comments improved this work.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.fsigen.2014.08.014.

References

[1] P. Gill, J.P. Whitaker, C. Flaxman, N. Brown, J.S. Buckleton, An investigation of the rigor of interpretation rules for STR's derived from less that 100 pg of DNA, Forensic Sci. Int. 112 (2000) 17–40.
[2] D.J. Balding, J. Buckleton, Interpreting low template DNA profiles, Forensic Sci. Int.: Genet. 4 (2009) 1–10.
[3] D. Balding, Evaluation of mixed-source, low-template dna profiles in forensic science, Proceedings of the National Academy of Sciences of USA. 110 (2013) 12241–12246.
[4] H. Haned, Forensim: an open-source initiative for the evaluation of statistical methods in forensic genetics, Forensic Sci. Int.: Genet. 5 (2011) 265–268.
[5] K. Lohmueller, N. Rudin, Calculating the weight of evidence in low-template forensic DNA casework, J. Forensic Sci. 58 (2013) 234–259.
[6] D. Taylor, J.-A. Bright, J. Buckleton, The interpretation of single source and mixed DNA profiles, Forensic Sci. Int.: Genet. 7 (2013) 516–528.
[7] M.W. Perlin, A. Sinelnikov, An information gap in DNA evidence interpretation, PLoS One 4 (2009) e8327.
[8] M.W. Perlin, M.M. Legler, C.E. Spencer, J.L. Smith, W.P. Allan, J.L. Belrose, et al., Validating TrueAllele® DNA mixture interpretation, J. Forensic Sci. 56 (2011) 1430–1447.
[9] B. Budowle, A.J. Onorato, T.F. Callaghan, A.D. Manna, A.M. Gross, R.A. Guerrieri, et al., Mixture interpretation: defining the relevant features for guidelines for the assessment of mixed dna profiles in forensic casework, J. Forensic Sci. 54 (2009) 810–821.
[10] C. Brenner, R. Fimmers, M.P. Baur, Likelihood ratios for mixed stains when the number of donors cannot be agreed, Int. J. Legal Med. 109 (1996) 218–219.
[11] S.L. Lauritzen, J. Mortera, Bounding the number of contributors to mixed DNA stains, Forensic Sci. Int. 130 (2002) 125–126.
[12] P.J. Green, Reversible Jump Markov Chin Monte Carlo Computation and Bayesian Model determination, Biometrika 82 (1995) 711–732.
[13] J.S. Buckleton, J.M. Curran, P. Gill, Towards understanding the effect of uncertainty in the number of contributors to DNA stains, FSI Genet. 1 (2007) 20–28.
[14] H. Haned, L. Pene, J.R. Lobry, A.B. Dufour, D. Pontier, Estimating the number of contributors to forensic DNA mixtures: does maximum likelihood perform better than maximum allele count? J. Forensic Sci. 56 (2011) 23–28.
[15] H. Haned, L. Pène, F. Sauvage, D. Pontier, The predictive value of the maximum likelihood estimator of the number of contributors to a DNA mixture, Forensic Sci. Int.: Genet. 5 (2011) 281–284.
[16] D.R. Paoletti, T.E. Doom, C.M. Krane, M.L. Raymer, D.E. Krane, Empirical analysis of the STR profiles resulting from conceptual mixtures, J. Forensic Sci. 50 (2005) 1361–1366.
[17] T. Tvedebrink, Overdispersion in allelic counts and θ-correction in forensic genetics, Forensic Sci. Int.: Genet. Suppl. Ser. 2 (2009) 455–457.
[18] J.-A. Bright, E. Huizing, L. Melia, J. Buckleton, Determination of the variables affecting mixed MiniFiler(TM) DNA profiles, Forensic Sci. Int.: Genet. 5 (2011) 381–385.
[19] J.-A. Bright, K. McManus, S. Harbison, P. Gill, J. Buckleton, A comparison of stochastic variation in mixed and unmixed casework and synthetic samples, Forensic Sci. Int.: Genet. 6 (2012) 180–184.
[20] J.-A. Bright, D. Taylor, J.M. Curran, J.S. Buckleton, Developing allelic and stutter peak height models for a continuous method of DNA interpretation, Forensic Sci. Int.: Genet. 7 (2013) 296–304.
[21] J.-A. Bright, J. Turkington, J. Buckleton, Examination of the variability in mixed DNA profile parameters for the Identifiler(TM) multiplex, Forensic Sci. Int.: Genet. 4 (2009) 111–114.
[22] S.J. Walsh, J.S. Buckleton, Autosomal microsatellite allele frequencies for a nationwide dataset from the Australian Caucasian sub-population, Forensic Sci. Int. 168 (2007) e47–e50.
[23] R.G. Cowell, T. Graversen, S.L. Lauritzen. Analysis of forensic DNA mixtures with artefacts, areXiv:13024404v2 [statME], 2013.

[24] D.A. Taylor, J.-A. Bright, J.S. Buckleton, The 'factor of two' issue in mixed DNA profiles, J. Theor. Biol. (2014), http://dx.doi.org/10.1016/j.jtbi.2014.08.021.
[25] M.D. Weinberg, I. Yoon, N. Katz, A remarkably simple and accurate method for computing the Bayes Factor from a Markov chain Monte Carlo Simulation of the posterior distribution in high dimension, areXiv:13013156v1 [astro-phIM], 14 January, 2013.
[26] M.D. Weinberg, Computing the Bayes factor from a Markov Chain Monte Carlo Simulation of the posterior distribution, Bayesian Anal. 7 (2012) 737–770.

Glossary

a: allele

c^2: variance constant used in the modelling of deviations of elements in **O** from elements in **E**, $\log\left(\frac{O_{ijkl}^2}{E_{ijkl}^2}\right) \sim N\left(0, \frac{c^2}{E_{ijkl}^2}\right)$

$\tilde{d}$: the distance chosen to encompass subset Ω_s around the mode of the entire posterior MCMC sample Ω

$d^{(y)}$: the distance of mass parameter values in posterior sample y from those in iteration y^*

D_n: the dimensionality of model n

E: vector of expected peak intensities

H_x: proposition x

K: the number of different models being examine i.e. the range of potential contributor numbers

l: locus

LR: the likelihood ratio

M: mass parameters; template amount for each contributor, degradation for each contributor, amplification efficiency for each locus, replicate amplification strength per replicate, stutter peak height variance and allele peak height variance

$\mathbf{M_n}$: mass parameters in model n

$\boldsymbol{M}_n^{(y*)}$: the set of mass parameter values that lead to the highest observed posterior likelihood in the MCMC sample Ω for model n

$M_{i,n}^{(y)}$: the value for mass parameter i at MCMC iteration y for model n (note the n can be dropped when talking about values all within a model)

n: number of contributors

N_n: model n

N: vector of all models under consideration

O: vector of observed peak intensities

P: the average value for a full profile genotype probability

r: replicate

R_r: replicate amplification efficiency for replicate r

$\mathbf{S_j}$: genotype set j

^{R}S: a genotype set that is present in non-zero weights 100% of the time for a specific contributor position

^{U}S: an unresolved genotype set, one where there are more than one possibility for a specific contributor position

$\boldsymbol{S_{U_j}}$: unknown, untyped contributor genotype set

$\mathbf{S_P}$: contributor genotype(s) known under H_1 but not H_2

$\mathbf{S_K}$: contributor genotypes known under both H_1 and H_2

$t_n^{(y)}$: the value for the template amount for contributor n in posterior sample y

V: the hyperrectangle volume

w_j: the weight for genotype set j

w_{jn}: the weight for genotype set j for model n

y^*: the iteration that lead to the highest observed posterior likelihood in the MCMC sample Ω

Z_n: a scalar for the weights in model n

$\Gamma(a, b)$: gamma distribution with shape of 'a' and a scale of 'b'

Ω: the entire posterior sample produced from the MCMC

$|\Omega|$: the number of individual samples in Ω

Ω_s: a subset of the posterior sample

$|\Omega_s|$: the number of individual samples in Ω_s

$\boldsymbol{\sigma}_q$: the vector of hyperrectangle edge lengths in iteration q of the algorithm outlined in supplementary material 2 section "Identifying Ω_s"

σ_{qp}: the hyperrectangle edge length in iteration q for mass parameter p

$\vee_{\boldsymbol{M_n},n}\, p(\boldsymbol{O})$: the maximum likelihood of the posterior MCMC sample Ω considering mass parameters $\boldsymbol{M}_n$ and model N_n

$\vee_y t_n^{(y)}$: the maximum observed template parameter value for contributor n in the y iterations of the posterior MCMC sample Ω

$\wedge_y t_n^{(y)}$: the minimum observed template parameter value for contributor n in the y iterations of the posterior MCMC sample Ω

Supplementary material 1

The *LR* corresponding to propositions where H_1 involves a POI who matches the resolved major contributor is likely to be largely unaffected by the addition of an unnecessary unknown contributor.

We demonstrate this by considering an *n* person profile, and examining the *LR* produced by comparison of a POI to the observed profile (***O***). As the fit to the profile will not substantially improve with increased numbers of contributors $Z_n \approx Z_{n'}$, but also weights between models will be the same (remembering that comparing between models means weights must be considered as likelihoods rather than residence times, as they are in [6]) simplifying equation 2 to:

$$LR = \frac{\sum_{j} \Pr(\boldsymbol{S}_j \mid H_1, N_n)}{\sum_{j'} \Pr(\boldsymbol{S}_{j'} \mid H_2, N_{n'})}$$

If some contributors have completely resolved genotypes that persist whether the number of contributors is *n* or *n'* then, in general terms, the *LR* considering resolved (*R*) and unresolved (*U*) contributors, as indicated by a left superscript, is:

$$LR = \frac{\Pr({}^{R}S \mid H_1, N_n) \sum_{k} \Pr({}^{U}\boldsymbol{S}_k \mid {}^{R}S, H_1, N_n)}{\Pr({}^{R}S \mid H_2, N_{n'}) \sum_{k'} \Pr({}^{U}\boldsymbol{S}_{k'} \mid {}^{R}S, H_2, N_{n'})}$$

In the circumstance where the POI matches the single resolved major contributor $\Pr({}^{R}S \mid H_1, N_n) = 1$, giving:

$$LR = \frac{\sum_{k} \Pr({}^{U}\boldsymbol{S}_k \mid {}^{R}S, H_1, N_n)}{\Pr({}^{R}S \mid H_2, N_{n'}) \sum_{k'} \Pr({}^{U}\boldsymbol{S}_{k'} \mid {}^{R}S_j, H_2, N_{n'})}$$

If we make the approximation that an average single contributor genotype has probability P then:

$$LR = \frac{\sum_{k} P^{n-1}}{P\sum_{k'} P^{n'-1}}$$

If the number of contributors under the two propositions is the same (i.e. $n = n', k = k'$) then the *LR* becomes:

$$LR \approx \frac{1}{P}$$

If $n' > n$ and the additional contributors are determined to add a very small contribution to the profile (as they are not required to explain the observed data) then $\sum_{k'} P^{n-1} = \sum_{k'} P^{n'-1} = 1$, as the genotypes that the additional contributor can take includes complete dropout (we designate a dropped out allele with 'Q') and cover all possibilities. This will again give:

$$LR \approx \frac{1}{P}$$

This suggests that the *LR* for a POI matching a single resolved major contributor is not significantly affected by the addition of unknown contributors under both H_1 and H_2 or just H_1 or just H_2.

Supplementary material 2

A1.1: Within model MCMC

We start first with a recap of the mathematics involved in within-model MCMC. We start with the *LR* from section 1.1:

$$LR = \frac{\Pr(\boldsymbol{O} \mid H_1)}{\Pr(\boldsymbol{O} \mid H_2)}$$

Introducing mass parameters $\boldsymbol{M}$ and genotype set, $\boldsymbol{S}$, and we obtain:

$$LR = \frac{\sum_j \int_M p(\boldsymbol{O} \mid \boldsymbol{S}_j, \boldsymbol{M}) p(\boldsymbol{M}) \Pr(\boldsymbol{S}_j \mid H_1) d\boldsymbol{M}}{\sum_{j'} \int_M p(\boldsymbol{O} \mid \boldsymbol{S}_{j'}, \boldsymbol{M}) p(\boldsymbol{M}) \Pr(\boldsymbol{S}_{j'} \mid H_2) d\boldsymbol{M}}$$

Note that we have dropped the H_x terms from all but the genotype set term, this is due to the independence of the other terms from the propositions. This independence from propositions allows the evaluation of the $p(\boldsymbol{O} \mid \boldsymbol{S}_j, \boldsymbol{M}) p(\boldsymbol{M})$ terms using MCMC and in particular the relationship:

$$p(\boldsymbol{M}, \boldsymbol{S}_j \mid \boldsymbol{O}) = \frac{p(\boldsymbol{O} \mid \boldsymbol{M}, \boldsymbol{S}_j) p(\boldsymbol{M}, \boldsymbol{S}_j)}{p(\boldsymbol{O})}$$

The analysis of $\boldsymbol{O}$ in the MCMC provides the posterior $p(\boldsymbol{M}, \boldsymbol{S}_j \mid \boldsymbol{O})$ and we require $p(\boldsymbol{O} \mid \boldsymbol{M}, \boldsymbol{S}_j) p(\boldsymbol{M}, \boldsymbol{S}_j)$ for use in the *LR*. The conversion of the $p(\boldsymbol{M}, \boldsymbol{S}_j \mid \boldsymbol{O})$ to $p(\boldsymbol{O} \mid \boldsymbol{M}, \boldsymbol{S}_j) p(\boldsymbol{M}, \boldsymbol{S}_j)$ can be achieved by multiplication by the marginal likelihood $p(\boldsymbol{O})$. This conversion is not carried out as within a model (single number of contributors) the

marginal likelihood will cancel in numerator and denominator of the *LR*. We are therefore taking advantage of the relationship $p(\boldsymbol{M},\boldsymbol{S}_j \mid \boldsymbol{O}) \propto p(\boldsymbol{O} \mid \boldsymbol{M},\boldsymbol{S}_j)p(\boldsymbol{M},\boldsymbol{S}_j)$.

Note that in [6] the residence time of each genotype set ($\boldsymbol{S}_j$) during the MCMC analysis is used as the weight, rather than the absolute value of $p(\boldsymbol{O} \mid \boldsymbol{M},\boldsymbol{S}_j)p(\boldsymbol{M})$ within the *LR*. This sum is a constant within a model and consistent in every term of the *LR*, so that, like $p(\boldsymbol{O})$ it cancels out and need not be enumerated. In effect using residence time has biased $p(\boldsymbol{O} \mid \boldsymbol{M},\boldsymbol{S}_j)p(\boldsymbol{M})$ by $\sum_j p(\boldsymbol{O} \mid \boldsymbol{M},\boldsymbol{S}_j)p(\boldsymbol{M})$. This is fine within a model, but between models the normalised values obtained using residence time must be adjusted back to likelihoods by multiplication by $\sum_j p(\boldsymbol{O} \mid \boldsymbol{M},\boldsymbol{S}_j)p(\boldsymbol{M})$. For the remainder of the supplement we do not mention this adjustment, as we assume the correction is made to the calculation to recover the absolute value of $p(\boldsymbol{O} \mid \boldsymbol{M},\boldsymbol{S}_j)p(\boldsymbol{M},\boldsymbol{S}_j)$.

We now introduce the model, $\boldsymbol{N}$ into the *LR* as another nuisance parameter and subscript dependent terms with the model term 'n':

$$LR = \frac{\sum_N \sum_j \int_M p(\boldsymbol{O} \mid \boldsymbol{S}_j, \boldsymbol{M}_{\boldsymbol{n}}, N_n)\, p(\boldsymbol{M}_{\boldsymbol{n}} \mid N_n) \Pr(N_n \mid H_1) \Pr(\boldsymbol{S}_j \mid H_1, N_n)\, d\boldsymbol{M}_{\boldsymbol{n}}}{\sum_N \sum_{j'} \int_M p(\boldsymbol{O} \mid \boldsymbol{S}_{j'}, \boldsymbol{M}_{\boldsymbol{n}'}, N_{n'})\, p(\boldsymbol{M}_{\boldsymbol{n}'} \mid N_{n'}) \Pr(N_{n'} \mid H_2) \Pr(\boldsymbol{S}_{j'} \mid H_2, N_{n'})\, d\boldsymbol{M}_{\boldsymbol{n}'}}$$

We again note that the only terms that are dependent on the propositions are priors for genotype sets and number of contributors. This again allows the evaluation of the remaining terms in the *LR* using MCMC. From Bayes theorem the model choice can be included in analysis of observed data by:

$$\frac{p(\boldsymbol{O}\,|\,\boldsymbol{M_n},\boldsymbol{S}_j,N_n)\,p(\boldsymbol{M_n},\boldsymbol{S}_j\,|\,N_n)}{p(\boldsymbol{O}\,|\,N_n)} = p(\boldsymbol{M_n},\boldsymbol{S}_j\,|\,\boldsymbol{O},N_n)$$

For the *LR* we require $p(\boldsymbol{O}\,|\,\boldsymbol{M_n},\boldsymbol{S}_j,N_n)\,p(\boldsymbol{M_n},\boldsymbol{S}_j\,|\,N_n)\Pr(N_n)$, by use of the relationship:

$$p(\boldsymbol{O}\,|\,N_n) = \frac{p\left(N_n|\boldsymbol{O}\right)p\left(\boldsymbol{O}\right)}{\Pr\left(N_n\right)}$$

We obtain:

$$\frac{p(\boldsymbol{O}\,|\,\boldsymbol{M_n},\boldsymbol{S}_j,N_n)\,p(\boldsymbol{M_n},\boldsymbol{S}_j\,|\,N_n)\Pr(N_n)}{p(N_n\,|\,\boldsymbol{O})\,p(\boldsymbol{O})} = p(\boldsymbol{M_n},\boldsymbol{S}_j\,|\,\boldsymbol{O},N_n)$$

Again the MCMC provides posterior sample $p(\boldsymbol{M_n},\boldsymbol{S}_j\,|\,\boldsymbol{O},N_n)$ which we multiply by $p(N_n\,|\,\boldsymbol{O})$. Note that again the $p(\boldsymbol{O})$ is omitted as now it is the probability of the observed data across all models being considered and so again will cancel out in the numerator and denominator of the *LR*. We again take advantage of proportionality, in this case $p(\boldsymbol{O}\,|\,\boldsymbol{M_n},\boldsymbol{S}_j,N_n)\,p(\boldsymbol{M_n},\boldsymbol{S}_j\,|\,N_n)\Pr(N_n) \propto p(\boldsymbol{M_n},\boldsymbol{S}_j\,|\,\boldsymbol{O},N_n)\,p(N_n\,|\,\boldsymbol{O})$.

The task is then to calculate the individual Bayes factors, $p(N_n\,|\,\boldsymbol{O})$, for comparisons between the *K* models.

A1.2: Defining the *Bayes factor*

We wish to consider *K* models, $\boldsymbol{N} = \{N_1,\ldots,N_K\}$ to assess some observed data $\boldsymbol{O}$. Each model, N_n, has a vector of mass parameters $\boldsymbol{M}_n \in \mathbb{R}^{\mathcal{D}_n}$ where $\mathcal{D}_n$ is the dimensionality of

the n^{th} model and n also serves as an indicator for the number of contributors, hence for example model N_2 would be a model considering two contributors to the DNA sample.

Bayes theorem gives the posterior probability density for each model as:

$$p(N_n|\boldsymbol{O}) = \frac{\Pr(N_n)\, p(\boldsymbol{O} \mid N_n)}{p(\boldsymbol{O})}$$

$\Pr(N_n)$ is the prior probability of model n and $p(\boldsymbol{O})$ is the unknown normalization constant. The marginal likelihood for model n is:

$$p(\boldsymbol{O} \mid N_n) = \int_{M_n} \sum_j p(\boldsymbol{O} \mid N_n, \boldsymbol{M_n}, \boldsymbol{S}_j)\, p(\boldsymbol{M_n} \mid N_n) \Pr(\boldsymbol{S}_j \mid N_n) d\boldsymbol{M}_n$$

The posterior odds of any two models $n_1, n_2 \in [1, K]$ are:

$$\frac{p(N_n|\boldsymbol{O})}{p(N_{n'}|\boldsymbol{O})} = \left[\frac{\Pr(N_n)}{\Pr(N_{n'})}\right]\left[\frac{p(\boldsymbol{O} \mid N_n)}{p(\boldsymbol{O} \mid N_{n'})}\right]$$

And is now independent of the normalisation $p(\boldsymbol{O})$. In the absence of propositions we choose the prior probability for all models to be equal, simplifying the above equation to:

$$\frac{p(N_n|\boldsymbol{O})}{p(N_{n'}|\boldsymbol{O})} = \frac{p(\boldsymbol{O} \mid N_n)}{p(\boldsymbol{O} \mid N_{n'})}$$

We concentrate now on the term $p(\boldsymbol{O} \mid N_n)$, known as the *Bayes factor*, which requires the calculation of the marginal likelihood for each model.

A1.3: Identifying the marginal likelihood

There are a variety of approximations for calculation of the marginal likelihoods, and one method commonly used (because of its simplicity) is the harmonic mean approximation (HMA):

$$p(\boldsymbol{O} \mid N_n) = \left[\frac{1}{|\Omega|} \sum_{y=1}^{|\Omega|} \frac{1}{p\left(\boldsymbol{O} \mid \boldsymbol{M}_n^{(y)}\right)} \right]^{-1}$$

Where Ω is the MCMC sample of the posterior probability and $|\Omega|$ is the number of samples within it. The HMA has the limitations that it is known to be highly variable [25, 26] and dominated by a few outlying low values for $p(\boldsymbol{O} \mid \boldsymbol{M}_n)$ [24]. This has led to the development of the different methods of assessing marginal likelihoods that avoid these issues [23, 24]. MCMC analyses of DNA profiles have the additional complexity that the genotype set represent a categorical parameter that, if consider as at the whole profile level, can contain an immense number of possible values.

Weinberg [23] makes the point that the marginal likelihood for model n, $P(\boldsymbol{O} \mid N_n)$, is defined by:

$$p(\boldsymbol{O} \mid N_n) \int_{\Omega_s} d \sum_j p(\boldsymbol{M}_n, \boldsymbol{S}_j \mid \boldsymbol{O}) = \int_{\Omega_s} \sum_j p(\boldsymbol{O} \mid N_n, \boldsymbol{M}_n, \boldsymbol{S}_j) p(\boldsymbol{M}_n, \boldsymbol{S}_j \mid N_n) d\boldsymbol{M}_n$$

Where the reduced set of the posterior sample $\Omega_s \subset \Omega$ is chosen to optimise the right hand integral above and reduce variability. We define $F(\Omega_s)$ as the fraction of points in Ω_s relative to Ω:

$$F(\Omega_s) = \frac{|\Omega_s|}{|\Omega|} = \int_{\Omega_s} d \sum_j p(\boldsymbol{M}_n, \boldsymbol{S}_j \mid \boldsymbol{O})$$

Giving:

$$p(\boldsymbol{O} \mid N_n) = F(\Omega_s)^{-1} \int_{\Omega_s} \sum_j p(\boldsymbol{O} \mid N_n, \boldsymbol{M}_n, \boldsymbol{S}_j) p(\boldsymbol{M}_n, \boldsymbol{S}_j \mid N_n) d\boldsymbol{M}_n$$

A1.4: Identifying Ω_s

We identify the subset of the posterior sample using the algorithm in section 3.1 of Weinberg 2013 [23]. This algorithm is transcribed below, but with terminology altered to be consistent with Taylor [6] . We consider a sample Ω containing y individual samplings from the posterior likelihood of an MCMC.

1) Select the parameter vector for model n, that corresponds to the maximum value of the posterior probability in the MCMC sample, $\bigvee_{\boldsymbol{M}_n, n} p(\boldsymbol{O})$. This corresponds to individual sample y^* and is vector $\boldsymbol{M}_n^{(y^*)}$. For simplicity we drop the model subscript for the remainder of this algorithm.
2) Calculate the shape of the hyperrectangle from the parameter range of the entire sample. If there are P parameters, $\boldsymbol{M} = M_1, \ldots, M_P$, then the hyperrectangle shape is defined by the vector:

   $\boldsymbol{\sigma}_0 = \left(max\{M_1\} - min\{M_1,\}, \ldots, max\{M_P\} - min\{M_P\} \right)$ and single element

   $\sigma_{0p} = max\{M_p\} - min\{M_p\}$
3) Compute the y distances from $\boldsymbol{M}^{(y^*)}$ by:

$$\left(d^{(y)}\right)^2 = \sum_{i=1}^{P}\left(M_i^{(y)} - M_i^{(y^*)}\right)^2 / \sigma_{qi}^2$$

Where q is the iteration of this algorithm and starts at $q=0$

4) Sort the y distances in ascending order and choose the distance value $\bar{d}$ so that a reasonable number of values have been chosen (e.g. $> 10{,}000$). The hyperrectangle shape that encloses this subset has coordinates for each parameter, p:

   $M_{p,min} = M_p^{(y^*)} - \sigma_{0p}\bar{d}$ and $M_{p,max} = M_p^{(y^*)} + \sigma_{0p}\bar{d}$

5) Increment q and recalculate the shape of the hyperrectangle from the variance for each parameter p of the entire sample by:

   $\sigma_{qp}^2 = \sum_{i=1}^{y}\left(M_p^{(i)} - M_p^{(y^*)}\right)^2$ including values where all parameters are enclosed within the previously defined hyperrectangle $\boldsymbol{M}^{(i)} \in [\boldsymbol{M}_{min}, \boldsymbol{M}_{max}]$

6) Step 4 – 6 can be repeated until the shape of the hyperrectangle has converged, however we do not do this

Note that the genotype set parameter $\boldsymbol{S}$ is categorical. For this parameter we sum across all genotype sets and so they are not defined within the hyperrectangle.

The hyperrectangle now encloses the desired subset Ω_s, which will contain the mass parameters for iteration y^*, girt by the other samples in Ω around the mode.

A1.5: Calculating the marginal likelihood

Weinberg found that $\int_{\Omega_s} \sum_j p(\boldsymbol{O} \mid N_n, \boldsymbol{M}_n, \boldsymbol{S}_j)\, p(\boldsymbol{M}_n, \boldsymbol{S}_j \mid N_n) d\boldsymbol{M}_n$ can be evaluated by identifying an *important region* around a dominant mode of the MCMC posterior sample, and using naïve Monte Carlo integration.

Naïve Monte Carlo integration is carried by:

1) Randomly sampling values for each of the parameters uniformly within the range enveloped by the hyperrectangle *Y'* times and recalculating the posterior likelihoods. At each iteration of the MC we calculate $\sum_j p(\boldsymbol{O} \mid N_n, \boldsymbol{M}_n^{(y')}, \boldsymbol{S}_j)\, p(\boldsymbol{M}_n^{(y')} \mid N_n, \boldsymbol{S}_j) \Pr(\boldsymbol{S}_j \mid N_n)$, using these parameter values (and using the original problem specification with the same prior and likelihood functions).
2) Calculating the average of the posterior likelihoods and multiplying by the volume, *V*, of the hyperrectangle

The hyperrectangle volume can be calculated by:

$$V = \prod_P \left(M_{p,max} - M_{p,min} \right)$$

The marginal likelihood is then calculated by:

$$p\left(\boldsymbol{O} \mid N_n\right) = \frac{V}{F\left(\Omega_s\right) Y'} \sum_{i=1}^{Y'} \sum_j p(\boldsymbol{O} \mid N_n, \boldsymbol{M}_n^{(i)}, \boldsymbol{S}_j)\, p(\boldsymbol{M}_n^{(i)} \mid N_n) \Pr(\boldsymbol{S}_j \mid N_n)$$

Note that we require genotype probabilities in the calculation of the marginal likelihood so that models can be appropriately weighted against one another. Strictly the calculation should be performed for each databases that is required for calculation of the *LR*, however it is

expected that the choice of population will have very little effect on the weights as profiles probabilities will generally be similar between populations.

Parameter vectors ***d*** (degradation) and ***t*** (template) change dimensionality with number of contributors. The hyperrectangle volume and $F(\Omega_s)$ terms will adjust between models of different dimensionality so that higher dimensional models are not unjustly penalized simply for having more parameters. Appropriate priors must be chosen for ***d*** and ***t*** when considering multi-dimensional problems. This is something which is not so important for within dimension problems where the uniform priors can be thought of as cancelling in an *LR* calculation, hence never requiring actual enumeration.

For multi-dimensional problems we specify priors as given below, using template DNA amount as an example.

Consider the prior for t_n as U[0,1], where 1 is the maximum value that template can take and 0 is the minimum value it can take. While the theoretical maximum value that template amount can take is infinite, a more useful range can be chosen for each contributor that encompasses some sensible range of values and depends on ***O***. Whilst this restriction isn't actually placed on the MCMC during its run, the sensible range can be thought of as the range of values visited by the MCMC during the analysis. So for template amount $\bigvee_y t_n^{(y)} = 1$ and $\bigwedge_y t_n^{(y)} = 0$. The hyperrectangle edge length for template then becomes the fraction of the range [0,1] that is encompassed in Ω_S.

A1.6: Shackling the hyperrectangle volume

We make one modification to the above method for determining the shape and position of the hyperrectangle, in order to minimise variability between runs without the need for impractically large numbers of MCMC iterations. Initially the above described process is carried out to determine the shape of the hyperrectangle. For all dimensional problems, other than the highest dimension, each hyperrectangle edge length is fractionally (and in equal relative proportions) changed so that all volumes between different dimensions are approximately equal to the volume of the hyperrectangle in the highest dimension. The $F(\Omega_S)$ then appropriately adjusts the marginal likelihood between dimensions.

A1.7: Definition of Z_n

We then define the 'weighting' for model *n* as:

$$p(N_n|\boldsymbol{O}) \propto Z_n = \frac{p(\boldsymbol{O}|N_n)}{\sum_K p(\boldsymbol{O}|N_n)}$$

So that in likelihood ratio calculations we have the intuitively pleasing (but mathematically unnecessary) result that $\sum Z_n = 1$.

A1.8: Incorporating Z_n into the *LR*

Starting with the *LR* from section A1:

$$LR = \frac{\sum_N \sum_j \int_M p(\boldsymbol{O} \mid \boldsymbol{S}_j, \boldsymbol{M}_n, N_n)\, p(\boldsymbol{M}_n \mid N_n) \Pr(N_n \mid H_1) \Pr(\boldsymbol{S}_j \mid H_1, N_n) d\boldsymbol{M}_n}{\sum_N \sum_{j'} \int_M p(\boldsymbol{O} \mid \boldsymbol{S}_{j'}, \boldsymbol{M}_{n'}, N_{n'})\, p(\boldsymbol{M}_{n'} \mid N_{n'}) \Pr(N_{n'} \mid H_2) \Pr(\boldsymbol{S}_{j'} \mid H_2, N_{n'}) d\boldsymbol{M}_{n'}}$$

And rearrange to:

$$LR = \frac{\sum_{N} \Pr(N_n \mid H_1) \sum_{j} \Pr(\boldsymbol{S}_j \mid H_1, N_n) \int_{M_n} p(\boldsymbol{O} \mid \boldsymbol{S}_j, \boldsymbol{M}_n, N_n)\, p(\boldsymbol{M}_n \mid N_n)\, d\boldsymbol{M}_n}{\sum_{N} \Pr(N_{n'} \mid H_2) \sum_{j'} \Pr(\boldsymbol{S}_{j'} \mid H_2, N_{n'}) \int_{M_{n'}} p(\boldsymbol{O} \mid \boldsymbol{S}_{j'}, \boldsymbol{M}_{n'}, N_{n'})\, p(\boldsymbol{M}_{n'} \mid N_{n'})\, d\boldsymbol{M}_{n'}}$$

The term $\int_{M_n} p(\boldsymbol{O} \mid \boldsymbol{S}_j, \boldsymbol{M}_n, N_n)\, p(\boldsymbol{M}_n \mid N_n)\, d\boldsymbol{M}_n$ is now weight for genotype sets and spans different numbers of contributors, and can be denoted w_{jn}. For conceptual ease w_{jn} can be thought of being the product of two weights, one within-model weight and one between-model weight. Defining

$w_{jn} = Z_n w_j = \int_{M_n} p(\boldsymbol{O} \mid \boldsymbol{M}_n, \boldsymbol{S}_j, N_n)\, p(\boldsymbol{M}_n \mid N_n)\, d\boldsymbol{M}_n \propto p(N_n \mid \boldsymbol{O})\, p(\boldsymbol{M}_n, \boldsymbol{S}_j \mid \boldsymbol{O}, N_n)$ we obtain:

$$LR = \frac{\sum_{n} Z_n \Pr(N_n \mid H_1) \sum_{j} w_j \Pr(\boldsymbol{S}_j \mid H_1, N_n)}{\sum_{n'} Z_{n'} \Pr(N_{n'} \mid H_2) \sum_{j'} w_{j'} \Pr(\boldsymbol{S}_{j'} \mid H_2, N_{n'})}$$

Which is Equation 2 from the text. As mentioned in section 1, *j* represents an exhaustive set of genotype sets, which are then incorporated into the *LR* with the propositions to produce *j* and *j'* non-zero element.

7.2 Clarification

Further description of the model in the paper supplementary

I have reused 'k' here as the summation index across unresolved contributors. This has caused confusion. I rewrite the derivation for supplementary material below:

It is useful to start with the general formula for the *LR*, given in the main body of the text:

$$LR = \frac{\sum_{n}\sum_{j=1}^{J} \Pr(\mathbf{O} \mid \mathbf{S_j}, N = n) \Pr(\mathbf{S_j} \mid Hp) \Pr(N = n \mid Hp)}{\sum_{n}\sum_{j=1}^{J} \Pr(\mathbf{O} \mid \mathbf{S_j}, N = n) \Pr(\mathbf{S_j} \mid Hd) \Pr(N = n \mid Hd)}$$

We then introduce a number of nuisance parameters, which we term mass parameters (**M**) that we integrate over:

$$LR = \frac{\sum_{n} \Pr(N = n \mid Hp) \sum_{j=1}^{J} \Pr(\mathbf{S_j} \mid Hp) \int p(\mathbf{O} \mid \mathbf{S_j}, N = n, \mathbf{M}) \Pr(\mathbf{M}) d\mathbf{M}}{\sum_{n} \Pr(N = n \mid Hd) \sum_{j=1}^{J} \Pr(\mathbf{S_j} \mid Hd) \int p(\mathbf{O} \mid \mathbf{S_j}, N = n, \mathbf{M}) \Pr(\mathbf{M}) d\mathbf{M}}$$

Let $w_j = \int p(\mathbf{O} \mid \mathbf{S_j}, N = n, \mathbf{M}) \Pr(\mathbf{M}) d\mathbf{M}$, which we term a weight. When likelihood ratios are calculated for a single number of contributors the weights are often displayed as normalised values so that they are more easily assessed by analysts. Let the normalising constant be Z_n, where:

$$Z_n = \sum_{j=1}^{J} w_j$$

So that the weights, normalised within a value of *n*, are:

$$w_j = Z_n w_{j,n}$$

The *LR* can then be written as:

$$LR = \frac{\sum_{n} Z_n \Pr(N = n \mid Hp) \sum_{j=1}^{J} w_{j,n} \Pr(\mathbf{S_j} \mid Hp)}{\sum_{n} Z_n \Pr(N = n \mid Hd) \sum_{j=1}^{J} w_{j,n} \Pr(\mathbf{S_j} \mid Hd)}$$

For simplicity of the following example, consider that the population model used assumes the probability of the genotype of a contributor is independent of the genotypes observed in other contributors (often referred to as the 'product' or 'Hardy-Weinberg' model). This allows the probability of a genotype set to be written as the products of the genotypes of the *n* individuals that make up the set, $\mathbf{S_j} = \left\{ {}^{1}G, ..., {}^{N}G \right\}$, so that:

$$\Pr(\mathbf{S_j} \mid Hp) = \prod_{i=1}^{n} \Pr({}^{i}G_j \mid Hp)$$

Again, for simplicity, consider a situation where the prosecution places all the probability on *n* contributors so that, $\Pr(N = n \mid Hp) = 1$ and defence place all their probability on *n'*, $\Pr(N = n' \mid Hd) = 1$. The *LR* is then:

$$LR = \frac{Z_n \sum_{j=1}^{J} w_{j,n} \prod_{i=1}^{n} \Pr({}^{i}G_j \mid Hp)}{Z_{n'} \sum_{j=1}^{J} w_{j,n'} \prod_{i=1}^{n'} \Pr({}^{i}G_j \mid Hd)}$$

Consider a situation where the DNA profile originates from *x* clearly resolved contributors, and that there is no indication in the profile (by way of peak height imbalance, drop-ins or other artefacts) of any more than *x* contributors of DNA. We can then split the genotype set probabilities into the resolved contributors and the remaining contributors needed to make the total up to *n* (under Hp) and *n'* (under Hd):

$$\prod_{i=1}^{n} \Pr({}^{i}G_j \mid H) = \prod_{i=1}^{x} \Pr({}^{i}G_j \mid H) \prod_{i=x+1}^{n} \Pr({}^{i}G_j \mid H)$$

Note that the alignment of contributors between *n* and *n'* i.e. contributor 1 in the set defined by Hp aligns with contributor 1 in the set defined by Hd, i.e. ${}^{i}G_j \mid Hp = {}^{i}G_j \mid Hd$. Within the sum across genotype sets the probabilities associated with the resolved contributors appear in every genotype set with non-zero value:

$$\sum_{j=1}^{J} w_{j,n} \prod_{i=1}^{n} \Pr({}^{i}G_j \mid H) = \left[\prod_{i=1}^{x} \Pr({}^{i}G_j \mid H) \right] \left[\sum_{j=1}^{J} w_{j,n} \prod_{i=x+1}^{n} \Pr({}^{i}G_j \mid H) \right]$$

Now consider that in this scenario where only *x* individuals are required to explain the evidence profile, that any additional contributors are assigned by the model to contribute approximately 0 fluorescence to the electropherogram. This has two consequences. Firstly, the weights associated with any of the *j* genotype sets will be approximately equal ($w_{j,n} \approx w_{j+1,n}$) so that:

$$\sum_{j=1}^{J} w_{j,n} \prod_{i=x+1}^{n} \Pr({}^{i}G_j \mid H) = w_n \sum_{j=1}^{J} \prod_{i=x+1}^{n} \Pr({}^{i}G_j \mid H)$$

Secondly, the *x* + 1 to *n* (or *n'*) contributors can possess any genotype at any locus. Therefore, the sum across all *J* possible genotype sets must equal 1 (as this is a sum of probabilities of genotype sets across all possible genotype sets that can exist):

$$\sum_{j=1}^{J} \prod_{i=x+1}^{n} \Pr({}^{i}G_j \mid H) = 1$$

So that the *LR* becomes:

$$LR \approx \frac{Z_n w_n \sum_{j=1}^{J} \prod_{i=1}^{x} \Pr({}^{i}G_j \mid Hp)}{Z_{n'} w_{n'} \sum_{j=1}^{J} \prod_{i=1}^{x} \Pr({}^{i}G_j \mid Hd)}$$

i.e. the probabilities associated with additional genotype sets of any additional contributors that are required to reasonably explain the profile have little to no impact on the *LR*.

Consider now that if there are only *x* completely resolved profiles (one for each contributor) that could describe the evidence profile that $J = 1$.

$$LR \approx \frac{Z_n w_n \prod_{i=1}^{x} \Pr({}^{i}G \mid Hp)}{Z_{n'} w_{n'} \prod_{i=1}^{x} \Pr({}^{i}G \mid Hd)}$$

The most common set of propositions considered in forensic genetics is that under the prosecution proposition one POI is nominated as a potential source of DNA, and that this person becomes an unknown in the defence proposition, i.e.:

Hp: The DNA originates from the POI and *n - 1* unknown individuals

Hd: The DNA originates from *n'* unknown individuals

Assume that we are consider a situation where the POI is not excluded from the profile i.e. the genotype of the POI (G_{POI}) aligns with one of the *x* resolved genotypes in the mixture (for simplicity let us say this is when $i = 1$ in the formulation above). Incorporating knowledge of the reference profile of the POI, and that fact that given Hp they are a contributor of DNA to the sample, so that:

$$\Pr({}^{1}G \mid G_{POI}, Hp) = 1$$

Yields an *LR* where the product over *x* contributors starts at element 2 in the numerator:

$$LR \approx \frac{Z_n w_n \prod_{i=2}^{x} \Pr({}^{i}G \mid Hp)}{Z_{n'} w_{n'} \prod_{i=1}^{x} \Pr({}^{i}G \mid Hd)}$$

If we consider that an average genotype probability is *P* then:

$$LR \approx \frac{Z_n w_n P^{x-1}}{Z_{n'} w_{n'} P^{x}} = \frac{Z_n w_n}{Z_{n'} w_{n'} P}$$

If the number of contributors is the same under Hp and Hd, $n = n'$, then the *LR* simplifies to:

$$LR = \frac{1}{P}$$

If $n \neq n'$ then the *LR* will depend on *P*, but also the probability associated with the two models, as given by the Z_n and w_n terms. Recall that $Z_n w_n$ represents the integration of the observed data over mass parameters:

$$Z_n w_{j,n} = w_j = \int p(\mathbf{O} \mid \mathbf{S_j}, N = n, \mathbf{M}) \Pr(\mathbf{M}) d\mathbf{M}$$

Within this integral the **M** term represents a number of parameters:

- Template DNA amount for each contributor (*n*), which has prior $t_n \sim U[0, T]$ (where T represents the upper limit on template amount before a DNA profile will no longer be analysed and is termed a saturation level)
- Degradation for each contributor, which has prior $d_n \sim U[0, D]$ (where *D* represents a level of degradation above which profiles will generally be considered too low quality and will not be analysed)
- A PCR replicate efficiency term for each PCR replicate (*y*), which has prior $R_y \sim U[0, \infty]$ (note that in practise, if an analysis was carried out and a replicate amplification efficiency obtained beyond the approximate bounds [0.1,10] it would be considered that one of the replicates is likely to have been the subject of an amplification error and should not be included in the analysis)
- An amplification efficiency term for each locus (*l*), which has prior $A^l \sim LN\left(0, \xi^2 \sigma^2\right)$ (where $\xi = \ln(10)$ is used to transform between logs in base 10 and base *e* and σ^2 is determined by laboratory calibration)
- A peak height variability parameter for each fluorescence type (*i*), which has prior $c^i \sim \Gamma\left(\alpha^i, \beta^i\right)$ (which is determined by laboratory calibration)

As priors for template and degradation (and also replicate amplification efficacy, but I do not include this parameter for reasons that will soon become apparent) are constants for all values of these parameters they can be taken outside the integral term. Let **M'** be the set of mass parameters without template or degradation so that:

$$Z_n w_{j,n} = w_j = (DT)^{-n} \int p(\mathbf{O} \mid \mathbf{S_j}, N = n, \mathbf{M}) \Pr(\mathbf{M}') d\mathbf{M}$$

For each contributor used to describe the evidence DNA profile, an additional template and degradation term are required. However, when these additional contributors do not contribute to the explanation of the profile then it is expected that:

$$\int p(\mathbf{O} \mid \mathbf{S_j}, N = n, \mathbf{M}) \Pr(\mathbf{M}') d\mathbf{M} \approx \int p(\mathbf{O} \mid \mathbf{S_j}, N = n+1, \mathbf{M}) \Pr(\mathbf{M}') d\mathbf{M}$$

And so the ratio of $Z_n w_n$ is

$$\frac{Z_n w_n}{Z_{n'} w_{n'}} = \frac{(DT)^{-n} \int p(\mathbf{O} \mid \mathbf{S_j}, N = n, \mathbf{M}) \Pr(\mathbf{M}') d\mathbf{M}}{(DT)^{-n'} \int p(\mathbf{O} \mid \mathbf{S_j}, N = n', \mathbf{M}) \Pr(\mathbf{M}') d\mathbf{M}} \approx (DT)^{n-n'}$$

In the running example this gives an approximate *LR* of:

$$LR \approx \frac{Z_n w_n P^{x-1}}{Z_{n'} w_{n'} P^x} \approx \frac{1}{P}(DT)^{n-n'}$$

In summary, the addition of unnecessary contributor(s) will only affect the *LR* by the additional prior probabilities incurred for template and degradation. There will therefore be a tendency to favour simpler models under these conditions for that reason, rather than a commonly held belief that the favouring of the simpler models would be due to smaller genotype set probabilities (resulting from an additional genotype in each genotype set, from the additional contributor).

7.3: The next generation of profiling technology and the need for next generation modelling

It was mentioned earlier that new technology promises to provide more data and from more STR regions. This technology is called Massively Parallel Sequencing (MPS), or sometimes Next Generation Sequencing (NGS). MPS has the ability to sequence multiple regions of DNA simultaneously, so that rather than just receiving size information (as is current DNA profiling) underlying sequences are also obtained. This technology targets a suite of different DNA markers that have different purposes. There are mutational markers (known as Single Nucleotide Polymorphisms, SNPs) that can be used to provide probabilistic assignments of physical features such as eye colour, hair colour, heights, age and appearance. There are also SNPs that are targeted to provide probabilistic assignments to populations of origin. Both of these marker types have strong investigative applications to unsolved crimes.

It is likely, however, that STRs will still be a main focus of identity based profiling for some time to come. This due to the fact that there are decades of legacy data (in the millions of profiles sitting in databases around the world) that require the same markers to be amplified if they are to be searched against. To this end, there is already much work being done in the field of sequencing STR markers using this type of new technology.

With STR data produced by MPS there will still be the need for probabilistic evaluation. While the underlying sequence will provide more discrimination power, there will still be unresolvable DNA mixtures to which individuals of interest will be compared. The new technology will have some similarities in the modelling of STR data as current size-based STR DNA profiling i.e. more DNA will lead to higher number of sequenced strands of that region (called 'coverage' in the parlance of MPS and akin to fluorescence in current STRmix™ modelling) and the DNA will still be degraded to varying degrees depending on the environment to which it has been exposed. There will also still be stutter and common alleles and stutters will still stack to produce a single indistinguishable data point. Loci will likely also amplify at different efficiencies.

However, new issues will arise. Sequencing errors will be a new factor to consider, and a model will have to be created that relates fragments of similar sequence to the possibility that they all arise from a common allele. There are also mechanics of the current MPS systems that appear to normalise the amount of DNA from each region within the sequencing reactions, which will affect the modelling of fluorescence currently employed. MPS techniques are also multi-step PCR reactions and this may require different modelling of peak height variability. Currently, there has not been enough work done to sufficiently model these factors and it is likely that it will take some time (years) before the same level of understanding is obtained for MPS derived STR profile behaviour as currently exists for current STR profile behaviour.

It is likely that as the studies are done and data starts to become available that models in existing continuous DNA interpretation system such as STRmix™ can be adapted to handle the new type of profiles.

Chapter 8: Discussion. Where to from here?

Chapter 8 places the work that has been described in the thesis up to this point in a wider context. The introduction of STRmix™ has largely addressed the biggest issue that forensic biology had, namely the ability to evaluate the complex DNA profile evidence being generated. There are (and always will be) a subset of results, or problem types, that cannot be addressed, but these are now very much in the minority compared to a decade ago.

The adoption of STRmix™ has been generally well received, both with forensic laboratories and by the legal community. The largest challenge faced by forensic analysts was the fear of the unknown, specifically the apprehension that they would not be able to understand the functioning of a complex statistical system (and therefore use and defend it). The biologists in the forensic community has underestimated their abilities, they have eagerly embraced the new knowledge and often provide insightful question or comments that direct further work and investigation.

From the legal community, the apprehension took a number of forms:

- that they would not be able to understand DNA evidence now
- that (specifically defence) could not challenge, nor find sufficient experts to challenge the new system
- that the testifying scientist could no longer be considered an expert
- of retrospective reanalysis of DNA results in old, closed cases

Each of these has been tested over the years via Court challenges. As a result of these challenges the rules of evidence law and the admissibility of scientific evidence in Australia have been changed in the High Court. In 2017, for the most part, the challenges have passed from the judicial system in Australia, and the forensic community have come to tolerate and embrace the new system of interpretation.

New issues have arisen in forensic biology, two areas being:

1) The interpretation of electrophoretic signal obtained prior to its evaluation as DNA profile evidence (and subsequent analysis in STRmix™)
2) The placement of the DNA evidence results obtained by using STRmix™ into a wider case context that is of interest to the court

Examples of each of these directions are given in the following two subsections of chapter 8.

8.1: Dealing with data pre-processing

Prior to analysis in STRmix™, prior to interpreting the data in an EPG to determine the number of contributors, someone (or more typically two independent people) have assigned each area of raised fluorescence (a peak) on the EPG as either requiring labelling or not. If a peak is representative of some modelled reason for the fluorescence, i.e. if it represents an allele present in the DNA samples, or is a stutter of the allele (if stutter modelling is present in the interpretation system) then the peak will be labelled. If it is a fluorescence caused by some mechanism that is not modelled by the interpretation system, for example pull-up, then it should not be labelled. During human interpretation of the EPG and the following STRmix™ analysis only the labelled information is considered. Whether the peak should be labelled will depend on its size, shape and position within the EPG. This description demonstrates that the process of labelling or unlabelling fluorescent data is:

a) Dependent on human judgement and so suffers from the same difficulties of any threshold or judgement based system
b) Highly important as it directly affects the downstream processes

Adding to the disadvantages of the current system is the fact that it is time consuming for two analysts to interpret EPGs, compare and resolve differences. Especially in modern hardware, such as the 3500xl capillary electrophoresis instruments, the dynamic range of fluorescence over which the instrument can function is so great that some peaks (from some contributors) can be highly intense, while others very low. The highly intense peaks cause artefacts to the DNA profile in the areas surrounding their position, both within the same dye lane and in other dye lanes. The analyst trying to interpret the fluorescent data must then distinguish the low-level artefacts from the low-level alleles.

A statistical tool that has come to recent popularity is artificial neural networks (ANN). Their particular strength is in pattern recognition and they have been used to great effect in this area, demonstrated by their ability to beat human in complex pattern recognition tasks as either part of a professional occupation or at game play. ANN therefore seem to be the perfect solution to the current problem of classifying fluorescence within an EPG. Further to this, ANNs work best when supplied with vast amounts of training data, and so a workflow could be imagined where a laboratory using ANNs to read DNA profiles, continually updates the training material and hence continually improves on the system in use. Such a model will act as an automatic updating system that works on a feedback loop to improve itself. This new area that lead to the publication in this section of the thesis.

Even if the system performs exceptionally (and experience so far is that it does) there will be challenges to its use. One of the early challenges will be the acceptance by the scientific and legal community. ANNs represent the ultimate in ‘mysterious statistics’ in that they are designed to take vast volume of data, teach themselves how to perform a task, often in ways that humans will not understand. There will have to be careful implementation of ANNs in a way that slowly integrates with current systems. This could be achieved in four stages:

1) Continue to use two individuals to read EPGs, but have the ANN as a tool that can suggest that currently labelled peaks may be artefactual. They are then being used simply as an assistant to the human reader, with the humans doing the ultimate decision making. Such a system would already greatly improve efficiency in DNA profile reading as it will make the reading process faster for the analysts, and is also likely to

result in less differences between two reading analysts. There would be the additional benefit of getting analysts used to the ANN and building some faith in their ability.

2) Stage 2 would occur once stage 1 had reached a point where analysts rarely overrode the ANN suggestion and when they had become comfortable with the ANN. Stage 2 would involve removing one of the readers, so that there was one human read and one ANN read that were compared. Again, this would further increase efficiency and consistency. In this stage that ANN use could migrates from a suggestion tool to the main peak detection and classification system.
3) Once stage 3 had progressed to the stage where the analysts were virtually never overriding the ANN peak classification the human reader could be removed. This would leave the ANN as the sole means of interpreting fluorescence on an electropherogram. The EPG would them be passed to analysts in the usual manner for human interpretation, prior to analysis in STRmix™
4) The final stage would require the integration of a few pieces of technology. Firstly, the ANN would read and classify fluorescence, which in the fully Bayesian manner would entail providing a probability for areas on the EPG being any one of the nominated categories (allele, stutter, baseline, pull-up, etc). This raw data would then be passed directly into a system like STRmix™ where the number of contributors would be treated as a nuisance variable and may become a parameter in the model (not requiring human pre-assessment). The deconvolution could then progress in an automated fashion. The result would be that no human interaction would occur until the stage of assessment of the STRmix™ analysis.

The result of the end of these four stages would be that instead of the analyst being handed a series of DNA profiles to read, interpret and analyse, they would be handed the completed package that they just needed to review. This would free up the analysts to consider the results in a wider case context (the focus of the next section).

Such a system is some way off, and may never be fully realised. The new DNA profile generation technology (as mentioned in section 7.3) is likely to drive the need for some level of automation in the manner described above.

Manuscript: Teaching artificial intelligence to read electropherograms. D Taylor, D Powers. (2016) Forensic Science International: Genetics 25, 10-18 – *uncited*

Statement of novelty: This work takes the theory of artificial neural networks and applies it to the problem of classifying fluorescence in a DNA profile into one of several, user-defined categories.

My contribution: Main author and sole simulation programmer. Equal contributor to theory.

Research Design / Data Collection / Writing and Editing = 50% / 100% / 80%

Additional comments:

Forensic Science International: Genetics 25 (2016) 10–18

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

Research paper

Teaching artificial intelligence to read electropherograms

Duncan Taylor[a,b,*], David Powers[b]

[a] Forensic Science South Australia, 21 Divett Place, Adelaide, SA 5000, Australia
[b] Flinders University, GPO Box 2100, Adelaide, SA 5001, Australia

ARTICLE INFO

Article history:
Received 5 June 2016
Received in revised form 24 July 2016
Accepted 26 July 2016
Available online 28 July 2016

Keywords:
Electropherogram
Gel reading
Artificial neural network
Artefact detection

ABSTRACT

Electropherograms are produced in great numbers in forensic DNA laboratories as part of everyday criminal casework. Before the results of these electropherograms can be used they must be scrutinised by analysts to determine what the identified data tells us about the underlying DNA sequences and what is purely an artefact of the DNA profiling process. A technique that lends itself well to such a task of classification in the face of vast amounts of data is the use of artificial neural networks. These networks, inspired by the workings of the human brain, have been increasingly successful in analysing large datasets, performing medical diagnoses, identifying handwriting, playing games, or recognising images. In this work we demonstrate the use of an artificial neural network which we train to 'read' electropherograms and show that it can generalise to unseen profiles.

1. Introduction

A common task for any forensic DNA laboratory is the generation of short tandem repeat (STR) DNA profiles. During polymerase chain reaction (PCR) fluorescently tagged primers are incorporated into copied DNA fragments. When run through a polyacrylamide filled capillary, the DNA amplicons are separated according to their molecular weight and ultimately passed through a laser. At this point electrons in the fluorophores are excited by the laser and, during de-excitation, emit light at particular wavelengths (depending on the fluorophore used) which are separated before being detected by a charge coupled device (CCD) camera. Carrying out this process produces a graph of detected light at several wavelengths over time (in seconds, or 'scans'), which we know as a DNA profile. Before these profiles can be used in interpretations they must be scrutinised by analysts to determine whether the information in the profile is representative of some component of DNA in the extract used to generate it, or if it is an artefactual product of the DNA profiling process. This task of 'reading' the electropherogram (EPG) can be time consuming and often leads to subjective differences between analysts. There have been a number of measures put in place to mitigate uncertainty in reading. The two most common are:

1. An analytical threshold (sometimes called baseline, or detection threshold) is employed, below which information will not be labelled. This works as the greatest level of artefactual fluorescence occurs at low intensities.
2. A system of double reading is commonly employed, whereby two analysts independently read the profile, and then compare their interpretations. Any differences are then sorted out and a consensus read is obtained. This helps to factor out reader differences.

Both of these measures have some drawbacks. The first measure requires a threshold be set at a value where the reading effort required by the analyst is practical for a high throughput laboratory. This is done at the expense of 'losing' any information that has not reached the threshold, which provides the counter-balance to the size of the set threshold.

The second measure doubles the level of resource required to read profiles, and there is no guarantee of consistency across different pairs of individuals.

1.1. Artefacts

The most common artefacts that are encountered in capillary gel electrophoresis can be grouped into two categories; stutters

* Corresponding author at: Forensic Science South Australia, 21 Divett Place, Adelaide, SA 5000, Australia.
E-mail address: Duncan.Taylor@sa.gov.au (D. Taylor).

http://dx.doi.org/10.1016/j.fsigen.2016.07.013
1872-4973/

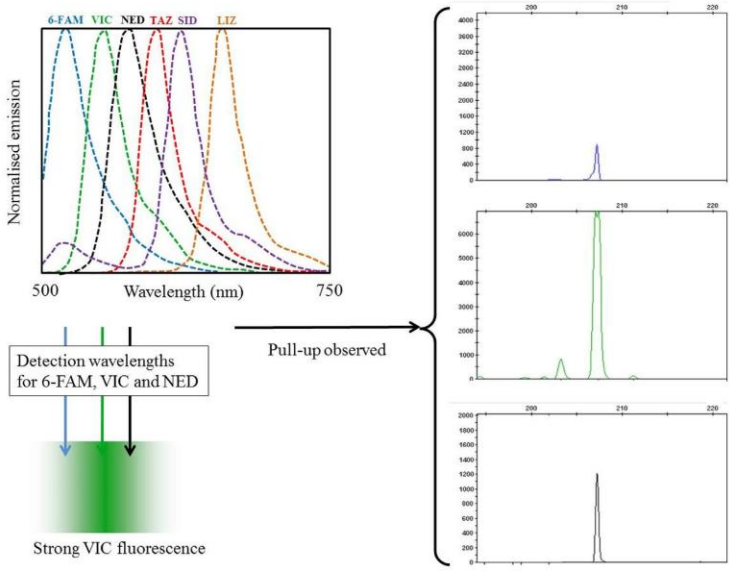


Fig. 1. Wavelengths of fluorophores used in the GlobalFiler™ PCR amplification kit (Life Technologies) showing an example of pull-up of a signal in the VIC (green) dye lane to the 6-FAM (blue) and NED (black) dye lanes. In the right hand electrophoretic graph the vertical axis designates intensity, in relative fluorescence units (RFU), and the horizontal axis represents the fragment size, in base pairs (bp). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

and pull-up.[1] Stutter occurs when there is an error during DNA replication of the STR during PCR. The result is that some copied fragments possess a different number of repeats from their parent template. Common stutters are one repeat shorter than their parent (referred to a back stutter, n – 4 stutter, or just stutter as it is the most commonly considered), one repeat longer (called plus stutter, forward stutter, n + 4 stutter or up stutter), two repeats shorter (double back stutter or n – 8 stutter) and half a repeat shorter (half stutter, or n – 2 stutter).

Pull-up occurs due to the overlap of the distribution of wavelength emitted by each of the fluorophores used in commercial profiling kits. The effect is that when a large number of fragments labelled with a specific fluorophore are detected by the CCD camera a high intensity peak is produced in the corresponding dye lane of the EPG and lower intensity peaks are seen in dye lanes that correspond to fluorophores with similar excitation wavelengths. This process is shown diagrammatically in Fig. 1. Other than the presence of peaks at a similar molecular weight, there are other features of pull-up peaks that can be used to distinguish them from allelic peaks. Two features are that pull-up peaks will not have any associated stutter peak, and the morphological characteristics of pull-up peaks (which are commonly described as 'spikey') tend to differ from the smooth Gaussian shape of allelic peaks.

It is the job of the analyst to recognise and remove artefacts such as pull-up that have not automatically been identified by an expert system. For stutters the situation is slightly different. Stutters require removal in reference samples as it is only the alleles that are of interest for most applications. For evidence samples the removal of stutter peaks will depend on the downstream system being used to interpret the profile. If this system is a continuous system (such as [1,2]) that utilises stutter peak in the modelling then the analyst will desire them to remain on the EPG. Otherwise they must be removed, and again there are automated ways in which this removal is carried out in expert systems, that identify most stutter peaks.

1.2. Expert EPG reading systems

There are two expert EPG reading systems that are in common use to read STR profiles, Genemapper (Life Technologies) and OSIRIS [3]. Both of these systems have some level of customisation that can be employed to automatically detect and remove artefacts in line with analysts' needs. Despite this customisation ability there remain a number of artefacts that still require manual removal by an analyst before the EPG can be used in criminal investigations. It is this manual removal of peaks that takes the most time during reading, and is one of the driving forces to increase analytical thresholds. In this paper we will discuss a novel method that could be incorporated into an expert system that utilises a technique known as artificial neural networks (ANN). Fig. 2 shows the subject of the experiment, which is a reference DNA profile that has been read in both Genemapper and OSIRIS. As the DNA profile is quite intense we expect a level of pull-up throughout the profile. In this work we will restrict the application of the ANN to just the 6-FAM dye lane (the blue trace) as this will demonstrate the power of the technique without overloading the reader with data. To this end, Fig. 3 shows the 6-FAM (blue) dye lane with peaks labelled as read in OSIRIS V2.5 (Fig. 3A), Genemapper ID-X (Fig. 3B) and then a contraction of the y-axis around the baseline to demonstrate the myriad of artefactual peaks

[1] There are a number of other artefacts that can be produced, such as dye-blobs, spikes or incomplete adenylation which are less common and we will not expand on these here.

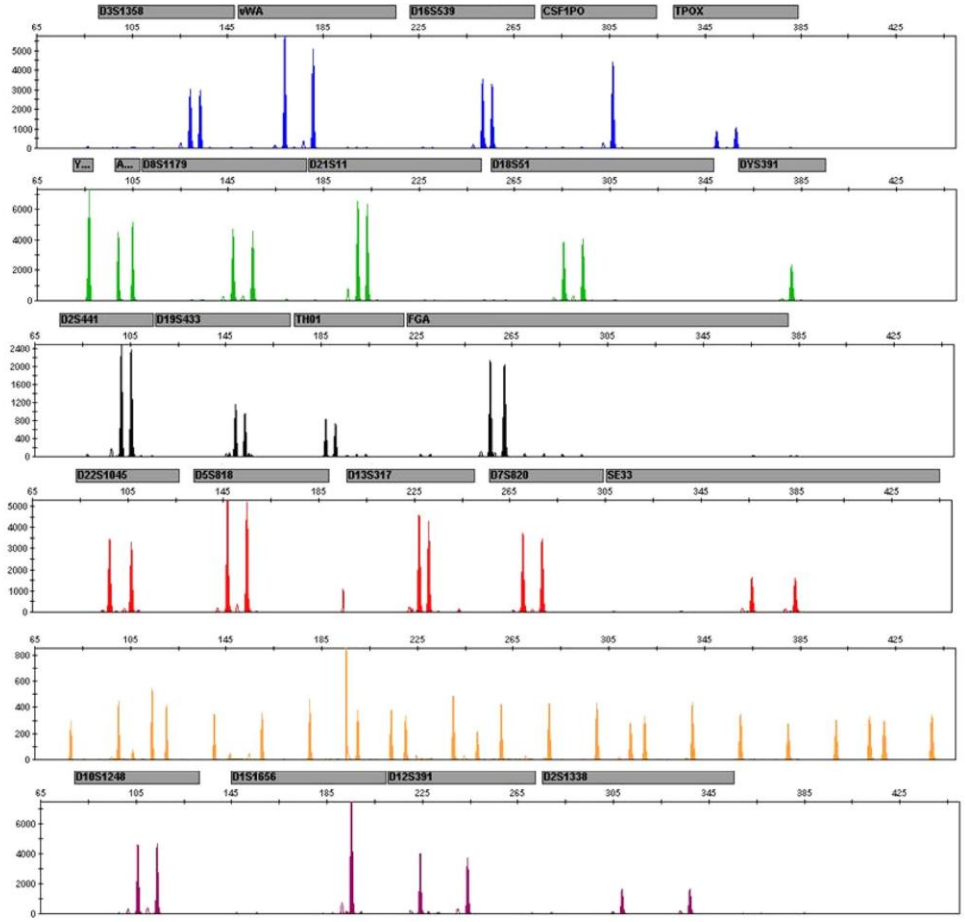


Fig. 2. DNA profile that will be the subject of the demonstration in this paper. The vertical axis is in RFU and the horizontal axis is in bp.

present in the profile. For the reading of the profile shown in Fig. 3 by both softwares, stutter filters have been applied which remove peaks deemed as stutter.

We note here that Fig. 3 is not meant as a comparison of software performance. Both have some level of user customisability and in particular OSIRIS has a number of advanced settings that deal specifically with the detection and removal of pull-up. These settings can be set to remove potential pull-up peak with some ferocity when reading reference samples, as the EPG is expected to possess only a limited number of high intensity peaks. When dealing with evidence samples the settings of programs to remove artefacts cannot be set at the same levels as the presumption of a limited number of similarly sized peaks cannot be made. In these instances a balance between correct artefact removal and incorrect allele removal must be struck, with most laboratories erring on the side of artefacts being labelled and needing to be assessed and removed manually by an analyst (or two).

Ideally an expert system would be able to use all the same features as humans to distinguish artefact from allele, with the advantages of:

- A consistent application of these features
- The ability to take numerous competing features into account simultaneously
- The ability to apply features of finer resolution than afforded by human ability

One potential avenue to achieve these goals is the use of ANN, which has shown success in a similar task of calling electrophoretic sequence data from either a summary of features obtained from pre-processing of electrophoretic data [4], or more recently the

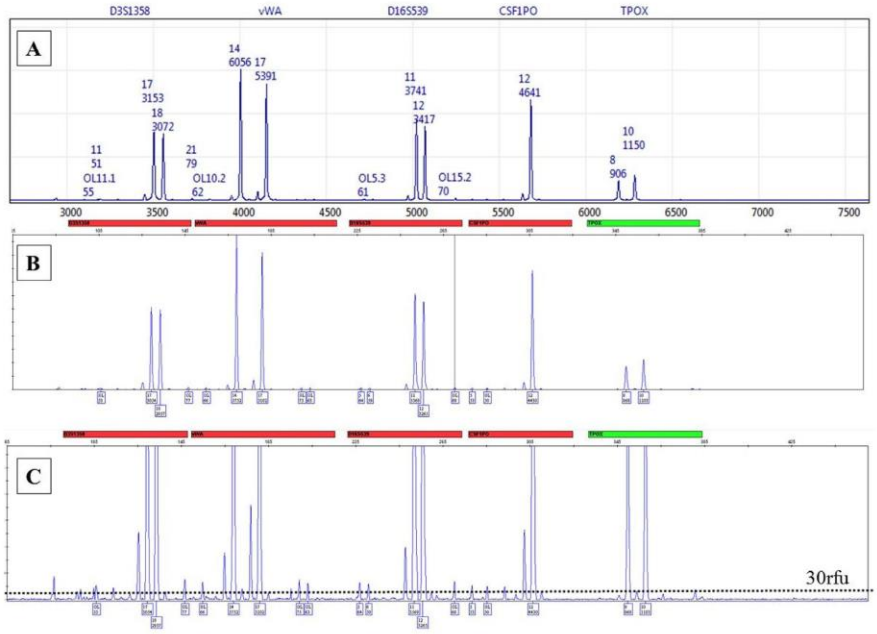


Fig. 3. 6-FAM dye lane of EPG shown in Fig. 2 as read (without user intervention) by OSIRIS (A) and Genemapper (B) and then a contraction of the y-axis around the baseline for the Genemapper read showing a typical analytical threshold of 30rfu (C). The vertical axis is in RFU and the horizontal axis is in bp.

training of electric current data produced from Nanopore sequencers [5]. ANN have also been used to identify the bounds of fluorescent bands in slab gel electrophoresis images [6].

1.3. Artificial neural networks

Artificial neural networks (ANNs) are a model of data processing that is designed to work like a human brain. The human brain possesses nearly 100 billion neurons [7], connected in a complex network of synapses. An individual neuron will possess a number of connections to other neurons and if the incoming signals from those connected neurons are received in a particular manner and strength then the neuron will activate, sending out signals of its own. This process is modelled within an artificial neural network, with the base unit of the model being called a neuron as shown in Fig. 4.

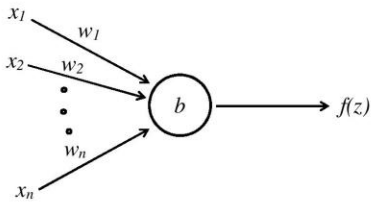


Fig. 4. A single neuron showing inputs (x), weights (w), bias (b) and output (f(z)).

In processing the inputs (x) the weighted (w) sum is added to the bias (b):

$$z = b + \sum_{i=1}^{n} x_i w_i \quad (1)$$

and the output is some function $f(z)$ (called the activation function), which can take a number of forms that perform in different ways. Real neurons are characterized as "firing" when some threshold of activation is reached, and information is often reflected in the rate of firing or encoded in the pattern of firing, but in common ANNs the activation function is typically a smoothed threshold function or "sigmoid".

In feedforward ANNs (FFNs) multiple neurons are used, arranged in layers as shown in Fig. 5. Inputs values are passed into a number of input neurons and after being propagated forwards through the hidden layers (hidden as the values are typically hidden, not being inputs or outputs) will result in a series of values in output neurons. If a 'softmax' layer is used in the final layer of an FFN then the output values are normalised and can be interpreted as probabilities.

In FFN shown in Fig. 5 there are only five inputs, two hidden layers and three outputs, but in practise learning neural networks have been trained that possess millions of neurons and over 100 billion parameters [8]. An FFN is also known as a multi-layer perceptron (MLP) and since a common supervised training method involves back-propagation of errors (BP) an MLP trained this way is called a back-propagation network.

Unlike classic regressions or modelling, a neural network is not specifically programmed to carry out a task in a particular way or using a specific model. Neural networks can carry out unsupervised or supervised learning. Unsupervised networks are given

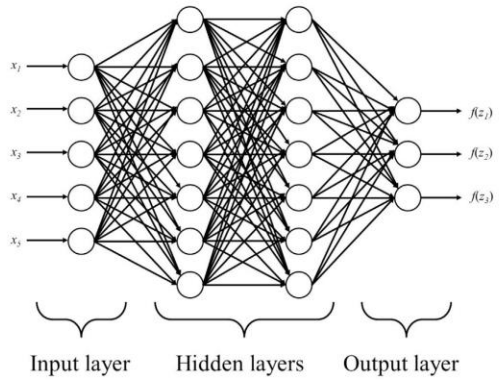


Fig. 5. Artificial neural network showing an input layer, two hidden layers and an output layer.

input data and seek to find patterns, or clusters containing common features, whereas supervised learning involves providing the neural network with a set of training data for which correct responses (labels) are also given, so that it can learn important features and generalise them for classification of unseen test data. Deep learning typically uses unsupervised learning of features on early layers and supervised learning of class labels in the later layers. We will be concentrating on supervised learning alone as this is the method we use for reading EPGs, but for the interested reader we recommend investigating ImagetNet (http://image-net.org) or Google DeepMind (https://deepmind.com) for an example of the power of artificial neural networks to analyse and classify data.

When supplied with training data the values are fed into input nodes, propagated forward through the network and result in values passed out of the output nodes, which are then compared to the supplied labels. This is carried out for J training instances and the cost for item j is determined by the loss (or cost) function:

$$L(W,B|j) = -\sum_{y\in O} y^{(j)} \ln\left[f\left(z^{(j)}\right)\right] + \left(1-y^{(j)}\right)\ln\left[1-f\left(z^{(j)}\right)\right] \qquad (2)$$

Eq. (2) shows the cross-entropy loss function, being one of a number of loss functions that can be used. In (2), O signifies all the neurons in the output layer, $y^{(j)}$ is the supplied response and $f\left(z^{(j)}\right)$ is the corresponding output. The total loss (or cost) is then obtained by averaging across all J input instances.

$$L(W,B) = \frac{1}{J}\sum_{j} L(W,B|j)$$

Training of the neural network is achieved by changing values of weights and biases to minimise the total loss through a process called gradient descent. The levels to which the weights and biases are changed depend on their contribution to the overall loss. This is determined through a series of partial derivatives of cost with respect to weight and biases (through intermediary 'error' values) that use the chain rule to move from outputs backwards through the network towards inputs in the aptly named process of back-propagation. We don't go through the mathematics of back-propagation as it is not required for conceptual understanding, but there are numerous texts available that do so. An accessible and comprehensive introduction to the topic is the free online textbook by Michael Nielsen (http://neuralnetworksanddeeplearning.com).

The final point we wish to raise is one of overfitting the model to the data. Neural networks can possess many parameters and the possibility exists for them to overfit to the training data i.e. learning aspects of that specific training data set rather than generalities or patterns. There are a number of ways that have been developed to avoid overfitting. One method, known as regularisation, adds a term to the loss function that means smaller weights (and hence a simpler model) is preferred. Another method is dropout, where for each training set some randomly chosen neurons are dropped from the network and the process of forward-propagation followed by back-propagation carried out. This prevents overfitting by effectively training over a series of different networks (working like an average across them) during training.

The most important factor however is to ensure strict separation of training and testing, including avoiding choosing parameters based on repeated tests with the test data. Often additional datasets called validation sets are used to allow for techniques like early stopping (stop if it starts getting worse on the validation set) or for tuning parameters. If there is not enough data available to create all these datasets, then techniques like bootstrapping or cross-validation are used that average over multiple allocations of samples to the different datasets. For example 10-fold cross-validation holds out 10% of the data for testing and trains on the other 90%, for all 10 test folds, and averaging over these results gives both a more reliable average performance estimate as well as an estimate of variance and opportunity to test significance [9,10]. In fact for our study, our first system performed very well, and we only explored closely variants of this system and did not find the need for cross-validation.

What remains to be divulged about neural networks is thus the myriad of ways in which they can be configured and constructed to best learn from the dataset being presented. Changes can be made to every aspect of neural networks that we have described, from the number of inputs and outputs, the number of hidden layers, the number of nodes in each hidden layer, the connection types, whether nodes are grouped into convolutional layers, the activation function used, the loss function used, the method of preventing overfitting, a series of complex network structure types that possess timed feed-back loops, and many others. The breadth of network types grows daily and for the interested reader Google is an excellent resource of information with many websites dedicated to ANN.

2. Method

The aim of this work was to demonstrate how an ANN could be trained to recognise different aspects of an EPG. The categories we chose were Baseline, Allele, Stutter, Pull-up and Forward Stutter. Ideally the trained ANN would be able to examine the 6-FAM dye lane in the EPG shown in Fig. 3 and classify each scan point as one of these five categories. In order to classify a particular scan, more than just the level of fluorescence of that 6-FAM scan point is important. Also important would be:

- Whether there is an area of fluorescence one repeat unit upstream in the 6-FAM lane
- Whether there is an area of fluorescence one repeat unit downstream in the 6-FAM lane
- Whether there is an area of fluorescence at the same scan range in other dye lanes

Therefore to classify each scan it was deemed that the input data would be the scan in question and 100 scans in either direction, in all dye lanes, which corresponds to approximately 8 base pairs (bp) in both directions under the conditions used to generate the EPG. This information is presented diagrammatically

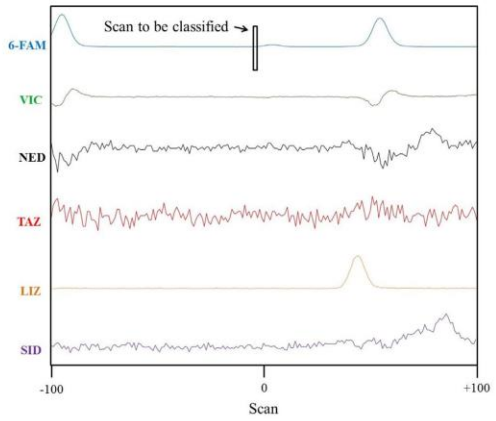


Fig. 6. Data used as input to classify central scan point in the 6-FAM dye lane as Baseline, Allele, Stutter, Pull-up or Forward Stutter.

in Fig. 6. The result is 201 scans in each of six dyes, leading to 1206 input neurons.

As there are five classification categories there are five output neurons in the neural network. The terminal layer was a softmax layer, meaning that the outputs could be interpreted as probabilities.

The remaining structure of the hidden layers in the ANN is then open for experimentation. For this demonstration, minimal optimisation was carried out in this regard, and there may well be ANN configurations that have higher correct classification rates, are simpler and require less input data. We started with a ANN similar to one that is known to produce good results recognising handwritten digits from the MNIST dataset [11]. This dataset is a common set used as a standard for neural network comparisons. It reads in 28×28 pixel images of digits, and hence has 784 input neurons, and classifies them into digit of 0–9, hence has nine output neurons. This is similar in size to the problem we are attempting. There are also similarities between the two datasets in that adjacent inputs are correlated and so a successful ANN for the MNIST classification task seemed like a logical starting point.

The ANN used in our study possessed one hidden layer of 100 neurons (as seen in Fig. 7). The cross-entropy loss function was used and the activation function was the rectifier function:

$$f(z) = \max(0, x) \text{ where } x \text{ is the input value}$$

which is known to work well for sparse data [12]. To combat overfitting dropout was employed with a rate of 0.2 for the input layer and 0.5 for the hidden layers.

The training data supplied were scans 3000–9000 for two reference profiles. Each scan within the 6-FAM dye lane was manually designated as Baseline, Allele, Stutter, Pull-up or Forward Stutter (shown in Fig. 8). No minimum level of signal was used at which data was unlabelled (or perhaps simply labelled as baseline), this means that some very minor perturbations of baseline signal have been designated as pull-up or stutter as seen in Fig. 8. We note that there is a level of subjectivity in this assignment, but the effect is likely to be minimal (explored further later on in the paper). Data before 3000 was not used as this is typically the zone where primer-dimer and unincorporated dyes come off the capillary and is an area of high signal that is ignored in EPG reading.

The result was 12,000 training sets of 1206 inputs. During training the entire 12,000 batch was run through the ANN for 100 iterations of training (referred to as 'epochs' in neural network parlance). The training process took five minutes using an Intel® Core™ i7-3940XM CPU @ 3.00 GHz with 32 GB RAM, running Windows 7 Ultimate.

Data was extracted from fsa files using the executable program fsa2xml supplied with OSIRIS and then manipulated to a csv file (with scan information listed in one column per dye format) using a custom written java program (available from the author on request). Raw scan data was scaled so that all values fell within the range [−1,1] with the mode of the baseline being 0. This was

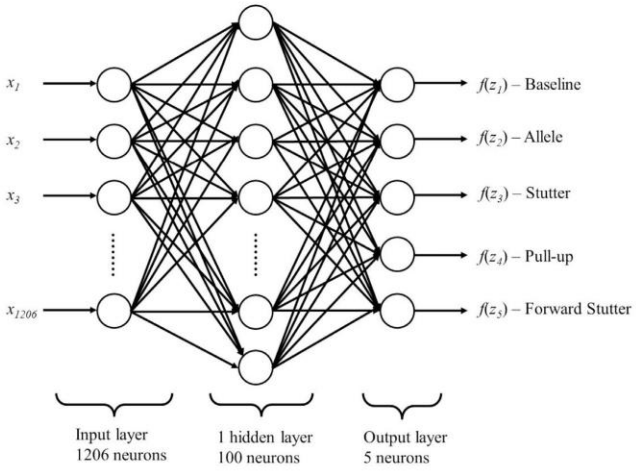


Fig. 7. ANN used in this study for classifying electrophoretic signal.

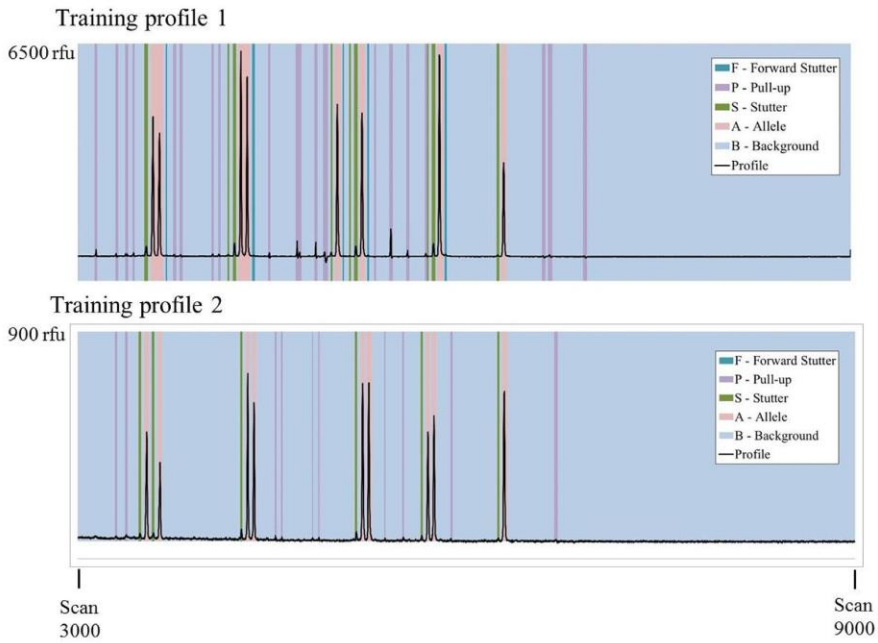


Fig. 8. Two 6-FAM training data sets used to train the ANN showing scans and approximate intensity in relative fluorescent units (rfu) as produced on a 3130xl (Life Technologies).

Table 1
Confusion matrix from trained ANN applied to 6-FAM dye data shown in Fig. 3. Rows represents ground truth responses and column are assigned classifications.

	Baseline	Allele	Stutter	Pull-up	Forward Stutter	Error rate (fraction)	
Baseline	4691	36	33	81	44	0.0397	(194/4885)
Allele	15	362	0	0	0	0.0398	(15/377)
Stutter	64	0	139	0	0	0.3153	(64/203)
Pull-up	97	0	9	333	7	0.2534	(113/446)
Forward Stutter	19	0	0	0	70	0.2135	(19/89)
Totals	4886	398	181	414	121	0.0675	(405/6000)

achieved for scan point i in dye d by:

$$scan^{d.i}_{transform} = \frac{scan^{d.i}_{original} - \text{mode}\left(scan^{d.}_{original}\right)}{10\,000}$$

where 10,000 is used as a divider because it is the theoretical maximum level of fluorescence detectedable by the 3130xl capillary gel electrophoresis instrument.

Table 2
Diagnostics for performance of ANN shown in Fig. 7. Total for recall and precision are weighted averages.

	Baseline	Allele	Stutter	Pull-up	Forward Stutter	Total
Recall	0.96	0.91	0.77	0.80	0.58	0.93
Precision	0.96	0.96	0.68	0.75	0.79	0.93
F score	0.96	0.93	0.72	0.77	0.67	0.80
G score	0.96	0.93	0.73	0.77	0.67	0.80
Informedness	0.79	0.91	0.76	0.78	0.58	0.79

To construct the ANN and carry out the learning, software R [13] was used with the H_2O addon (http://www.h2o.ai/). All R code for transforming data, creating training and test datasets and building and training the ANN has been supplied as Supplementary material.

3. Results

On the training dataset the ANN was able to learn to correctly classify approximately 98% of the 12,000 scans. When the model was then applied to the test dataset the performance was slightly lower at approximately 93%, but still generally high (see confusion matrix results in Table 1).

Table 2 shows the greatest level in confusion was assigning scans into the Baseline classification. In many instances this will have little practical effect i.e. whether a low intensity area of fluorescence is low level pull-up or slightly raised baseline noise will not affect reading as neither of these categories of data is going to be labelled during reading. When scans marked manually as

Allele were assigned into the Baseline classification (and vice versa) it was at the tails of the allelic fluorescence, where it drops back to baseline levels. This is likely due to the relatively arbitrary choice of where along this tailing curve we chose to switch from classifying scans from Allele to Baseline and so is again of little practical consequence. The highest percentage of misclassification occurred with the Forward Stutter classification and this is likely a product of relatively little training data (the dark blue category shown in Fig. 8). It is likely that a larger training dataset would alleviate this confusion.

To visually represent the classification of the trained ANN on the 6-FAM scans, barplots were produced for the probability of that scan being each of the classifications and this is shown in Fig. 9.

Of particular note is the ability of the ANN to distinguish between the five categories after training with only two example profiles. The training data (whilst somewhat mundane to produce) is virtually limitless and expanding the training set would no doubt improve performance.

To assess the performance of the ANN we consider informedness, a measure of the power of the model to predict outcomes better than chance. Formally, informedness is a measure of how informed a predictor is for the specified condition, and specifies the probability that a prediction is informed in relation to the condition (versus chance) [14,15]. The informedness of the ANN, from the data in the confusion matrix (Table 1) was 0.79. Details of the diagnostics are given in Table 2.

4. Discussion

There is real promise in implementing ANNs into expert EPG reading systems. The data type, with clear features and patterns, makes it ideal for processing in this manner. Fig. 8 shows that in general, allelic data was correctly classified and there were no instances where artefactual peaks were classified as allelic. If this information was carried forward to an expert system such as OSIRIS or Genemapper then the instruction to the software would be to only label peaks that fall within the ranges classified as allelic (or stutter if the downstream interpretation system required it). This would avoid the multiple manual removals that are currently required by an analyst (as seen in Fig. 3) and ultimately save many hours of reading. The eventual goal would be to have an ANN that had been trained sufficiently that no human intervention was required at all, a goal that does not seem unrealistic given the many diverse demonstrated achievements of ANNs.

The current study has a number of obvious areas of expansion before it neared the goal of being used, unchecked, in an expert

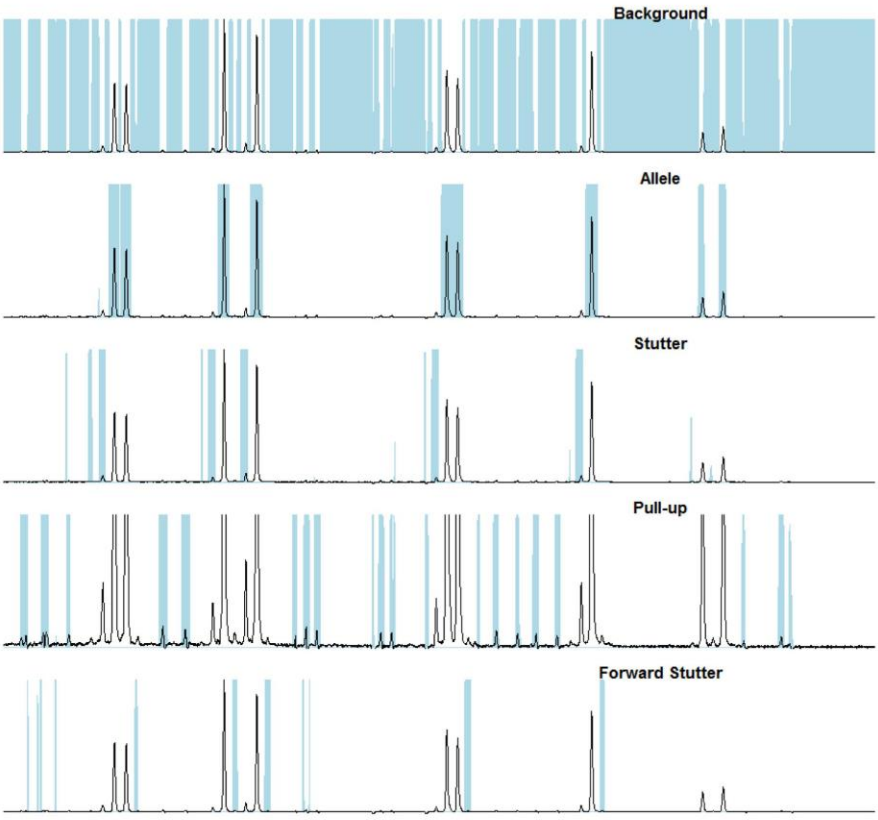


Fig. 9. Classification probability (shaded bars) as determined by the trained ANN when applied to the test data. The vertical axis contracted to the baseline for the pull-up category to show performance.

system. The demonstration in this paper has chosen only a single dye lane in arguably the simplest gel reading situation that could have been chosen i.e. single source profiles of reasonable quality. Further training and testing would initially need to include references that are much weaker in intensity, or perhaps complicated by issues such as incomplete peak resolution, trisomies or complex stutter patterns. Not all loci possess STRs that have a 4 bp repeat motif, some loci (such as Amelogenin and Yindel) are not STR loci and so have no stutter, other loci have repeat motifs of three or five bp. To account for these differences may necessitate separately trained neural networks for each locus, or a dataset that includes additional inputs that describe these known features. Alternatively a larger set of classifications could be used that classify the central scan of all dye lanes simultaneously (but given the vast increase in classification categories this would need a much larger training set). The next goal would then be to apply the ANN model to complex mixed EPGs. It is these profiles where the most time is currently spent by analysts interpreting the fluorescent signals.

The demonstration in this paper uses a relatively (by ANN standards) small training dataset, which could easily be expanded by the addition of profiles that are produced routinely every day in typical forensic laboratories. Each addition of training material will take less time to prepare as the currently trained ANN can be used to classify each scan in the next prospective training addition and will likely automatically correctly classify the vast majority of scans, only requiring minor modification by the analyst (most likely when some new data feature is being added to the training material). It may also be possible to reduce the number of inputs from what has been used here. One potential reduction would be the areas of outside approximately ± 10 scans in all dye lanes other than the one being labelled (as the fluorescence in other dye lanes is really only useful to identify pullup which occurs during the same scans).

4.1. World without an analytical threshold

The analysis carried out in this work has been done so without truncating the input data at any particular threshold. This brings up the interesting possibility that if a ANN was used to classify scans as allelic or artefactual that there would not need to be any analytical threshold employed for interpretation. All data could be assessed truly without thresholds. This is quite different to the current ubiquitous methodology of implementing such a threshold when reading electrophoretic data. The use of the outputs of a ANN with a softmax layer could be used to probabilistically label peaks, e.g. label any area of fluorescence which has a probability of greater than 0.99 (for example) of being allelic. As allelic data reduces in intensity, until it was eventually subsumed into the baseline noise, the probability of it being allelic (according to the assignment of the ANN) would drop until it fell below the probabilistic cutoff and was no longer labelled. Peak morphology would also be taken into account in this assessment as the ANN learnt higher level features of what constitutes fluorescence from allelic fragments. In such a world the classic definition of dropout would have to change as it is commonly referred to as the probability of a peak, expected at some intensity, falling below the analytical threshold. Instead it may need to be defined as the probability of too few strands of DNA being sampled from a DNA extract to produce fluorescence that will be recognised as allelic. Better still, the probabilities associated with all scans being allelic, baseline, stutter or pull-up could be fed directly into an expert interpretation system, which would utilise them when determining potential contributing genotypes. There is much potential for expansion in this area.

5. Conclusion

Artificial neural networks are a tool that is becoming more and more popular to find patterns in, and make predictions from, large amounts of data. Electrophoretic signals that make up EPGs are a perfect candidate for applying ANN to reduce the subjective and laborious task of manually classifying data as allelic or artefactual. In this work we have demonstrated a simple ANN that, when trained on only two EPGs, was then able to identify areas of allelic fluorescence completely in the next EPG it was given.

Much work is required to develop and train a ANN that could be used routinely in active forensic casework; however the advantages of pursuing such a system are great. Not only would it save resources, it will allow access to data that is currently lost due to the application of an analytical threshold. Although not shown here it is also possible to probe the weights and biases of trained ANN to determine which features of the inputs are the most important in classifying data. It may well be that ANN could identify and teach us about features of the data we are producing that we could not see otherwise.

Acknowledgement

Points of view in this document are those of the author and do not necessarily represent the official position or policies of their organisations.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.fsigen.2016.07.013.

References

[1] M.W. Perlin, M.M. Legler, C.E. Spencer, J.L. Smith, W.P. Allan, J.L. Belrose, et al., Validating TrueAllele® DNA mixture interpretation, J. Forensic Sci. 56 (2011) 1430–1447.
[2] D. Taylor, J.-A. Bright, J. Buckleton, The interpretation of single source and mixed DNA profiles, Forensic Sci. Int. Genet. 7 (2013) 516–528.
[3] R.M. Goor, L.F. Neall, D. Hoffman, S.T. Sherry, A mathematical approach to the analysis of multiplex DNA profiles, Bull. Math. Biol. 73 (2010) 1909–1931.
[4] O. Mohammed, K.T. Assaleh, G.A. Husseini, A.F. Majdalawieh, S.R. Woodward, Novel algorithms for accurate DNA base-calling, J. Biomed. Sci. Eng. 6 (2013) 165–174.
[5] V. Boža, B. Brejová, T. Vinař, Deepnano: Deep Recurrent Neural Networks for Base Calling in Minion Nanopore Reads, ArXiv: 160309195v1 [q-bioGN], 2016.
[6] M.K. Turan, A. Elen, E. Sehirli, Analysis of DNA gel electrophoresis images with backpropogation neural network based Canny edge detection algorithm, Int. J. Sci. Technol. Res. 2 (2016) 55–63.
[7] F. Azevedo, L. Carvalho, L. Grinberg, J.M. Farfel, R. Ferretti, R. Leite, et al., Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain, J. Comp. Neurol. 513 (2009) 532–541.
[8] A. Trask, D. Gilmore, M. Russell, Modeling Order in Neural Word Embedding at Scale, arXiv: 150602338v3 [csCL], 2015.
[9] D. Powers, A. Atyabi, The problem of Cross-Validation: Averaging and Bias, Repetition and Significance Engineering and Technology (S-CET), Spring congress on. 2012, IEEE, 1–5.
[10] I. Witten, E. Frank, M. Hall, Credibility: Evaluating What's Been Learned Data Mining Practical Machine Learning Tools and Techniques, Elsevier Inc., Burlington, USA, 2011 Chapter 5.
[11] Y. LeCunn, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (1998) 2278–2344.
[12] X. Glorot, A. Bordes, Y. Bengio, Deep sparse rectifier neural networks, Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS-11) (2011) 315–323.
[13] M. Plummer, Bayesian graphical models using MCMC, 2012. rjags.
[14] D. Powers, Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation, J. Mach. Learn. Technol. 2 (2011) 37–63.
[15] D. Powers, Evaluation evaluation a monte carlo study, European Conference on Artifical Intelligence (2008) ArXiv: 150400854 [csAI].

8.2: Placing the statistical DNA profile evaluations within a wider case context

Over the years, that the questions being asked in court regarding DNA evidence have shifted from "*Whose DNA is this?*" to "*How did it get there?*". In the parlance of the hierarchy of propositions (see section 3.2 for an explanation of this concept) the questions were activity level rather than being source (or sub-source, or sub-sub-source) level. STRmix™ answers questions at sub-source level and so there was a disconnect between the information being provided and the questions being asked.

There is a common misunderstanding that if the questions are about activity that the sub-source level work is no longer relevant, hence sweeping away the need for systems such as STRmix™. This is not the case. Moving up through the hierarchy of propositions is like building house, which must be based on solid foundations. STRmix™ has provided that foundation, and it is only through the existence of software that can address the sub-source level propositions, that it becomes possible to consider higher-level propositions.

There are a number of publications that explain the evaluation of forensic findings to help address activity level propositions. These include biology-focussed publications. There is a movement beginning within Australian forensic biology laboratories to develop the ability to numerically assess findings considering activity level propositions to forensic biology in Australia. This alone is a good driving force to conduct research in this area, although not the motivation behind the published work provided below. The case involves an alleged attempted abduction that occurred locally, and for which FSSA analysed key items from both the alleged attacker and the alleged victim. There were no signs of DNA from either party on the other's clothing and the DNA analyst was called to testify. In these sorts of cases (i.e. cases where there is no DNA support for the allegation) it is not unusual to still be called by the prosecutor, who wishes to make the point that just because there was no DNA detected, it doesn't mean the activity didn't take place. The line of reasoning is a rewording of the old adage "*absence of evidence is not evidence of absence*", but in reality, absence of evidence is indeed evidence of absence (in that the lack of detected DNA will tend to support a scenario of non-contact over contact), just not conclusive evidence. So too was the prosecutor's intent in this case and the defendant was ultimately convicted of the alleged crime (presumably based on non-DNA evidence).

Later, on review of the results, it was deemed by the court of appeal that the DNA testimony was misleading and the conviction was overturned. This caused some concern locally, as the analyst had not misrepresented the DNA findings, and there were dozens of similar cases that had been testified to in a similar manner. Work was conducted to put the DNA results in a wider case context (in this particular case the DNA results showing an exclusionary result) by carrying out a full activity level (and in the paper even went through an offense level) analysis was performed. The aims in doing so were three-fold:

1) To show that the court of appeal had unfairly judged the testimony of the scientist and that the issue was a wider misunderstanding of the levels in the hierarchy of propositions.
2) To demonstrate how the data could be evaluated to help address the questions of interest. In particular, it was hoped that this work could be a useful example to point to when faced with the line of questioning "*just because there was no DNA detected, it doesn't mean the activity didn't take place*"

3) To promote the practise of considering DNA findings in a wider case context

Manuscript: The evaluation of exclusionary DNA results: a discussion of issues in R v. Drummond. D Taylor. (2016) Law, Probability and Risk 15 (3), 175-197 – *uncited*

Statement of novelty: This work applies the laws of probability and utilises Bayesian Networks to demonstrate the benefits of considering activity level propositions when evaluating the DNA results in a wider case context. In particular, this work demonstrates the importance of evaluating an exclusionary result.

My contribution: Sole author.

Additional comments:

Law, Probability and Risk Advance Access published July 11, 2016

Law, Probability and Risk (2016) **0**, 1–23 doi:10.1093/lpr/mgw004

The evaluation of exclusionary DNA results: a discussion of issues in *R* v. *Drummond*

DUNCAN TAYLOR†

Forensic Science South Australia, 21 Divett Place, Adelaide, SA 5000, Australia; School of Biological Sciences, Flinders University, GPO Box 2100 Adelaide SA, Australia 5001

[Received on 27 September 2015; revised on 9 May 2016; accepted on 10 May 2016]

A recent appeal ruling by the Supreme Court of South Australia, based in large part on the significance of the absence of an individual's DNA on an item, has been successful in overturning a conviction. An issue of interest to the court in the original trial was the probability that two people could struggle and DNA not be detected on each other's clothing. In the parlance of the hierarchy of propositions, the question of interest to the court was about potential activities, however using the DNA results the expert could only provide source level information and in the absence of DNA results such reporting is meaningless. In this article the circumstances of the case, the evidence at trial and the expert reports are examined to understand the circumstances that led to the initial conviction and subsequent successful appeal. The main body of this article is dedicated to helping evaluate the DNA profiling results considering competing posited activities to determine the value of the exclusionary evidence to the questions posed in court. I demonstrate the application of Bayesian theory and relevant literature studies on DNA transfer to assign a likelihood ratio, which in this case supported the defence version of events.

Keywords: Hierarchy of propositions; Drummond; activity; likelihood ratio; appeal; transfer; persistence.

1. Introduction

A recent appeal ruling by the Supreme Court of South Australia *R v Drummond* (2015) has overturned an earlier conviction based in large part on the significance of the absence of an individual's DNA on an item. There are a number of important concepts that contribute to the discussion of this matter. I explore here the circumstances of the case, the forensic evidence and the reasons behind the ultimate quashing of the conviction. I also explore the value of the evidence using data from relevant published studies and considering the questions posed about disputed activities during the trial.

There are a number of details of the case that I will omit from my explanation of the scenario. I will concentrate only on those exhibits and results that relate to the subject of the appeal. There were other exhibits seized and many aspects of evidence, both expert and eye witness, that I do not mention.

1.1 *The circumstances of the case*

The prosecution case stated that the victim in this matter (T) was walking along Prospect Road on 24 November 2010 wearing a singlet top. It is alleged that a car pulled over, Drummond exited, grabbed T by the arm and attempted to pull her into the vehicle. T struggled with Drummond, which included her

†Corresponding author. Email: Duncan.Taylor@sa.gov.au

hitting him in the chest and he abandoned his attempt to kidnap T and drove off. T noted the licence plate of the vehicle.

T called police stating that she had been attacked and provided the licence plate. Shortly after police attended the home of Drummond and seized his clothing that he had been wearing at the time in question. The following day police seized as evidence the clothes worn by T.

Drummond stated that he was driving home along Prospect Road, but never stopped his vehicle. There was an implication that T had not been attacked at all but this was not explored in any detail. The main issue in court appeared to be the identity of T's attacker.

1.2 *The forensic results*

At the forensic laboratory the tops of Drummond and T were examined and tapelifts taken from each. From a tapelift of the singlet top of T a mixed DNA profile was obtained that could be explained by the presence of Deoxyribonucleic acid (DNA) from three individuals. T's alleles corresponded to the major component of the mixture and Drummond was excluded as a minor contributor by the forensic scientist. A male friend of T, who had previously hugged her that day, could account for the alleles of one of the minor contributors. Y-STR profiling was conducted and a single sourced profile containing the same alleles as T's male friend was obtained. No sign of a second male was detected from Y-STR analyses and so the conclusion was that the third component of the mixture from T's top was female.[1]

From a tapelift of the top of Drummond a mixed DNA profile was obtained that could be explained by the presence of DNA from two individuals. Drummond's alleles corresponded to the major component of the mixture and T was excluded as the minor contributor by the forensic scientist.

DNA extractions were carried out using DNA-IQ[2] (Promega). DNA profiling was carried out using Profiler Plus™[3] (Life Technologies) and interpretations were carried out manually (i.e. not assisted by computer software) using the thresholds and guidelines appropriate in 2010 at Forensic Science SA (FSSA).

1.3 *Evidence at trial by the forensic scientist*

Evidence was given at the trial on the absence of T's reference in the profile obtained from the top of Drummond and Drummond from the top of T. The questioning turned to the chance of an individual not leaving detectable levels of DNA on another's clothing from the above described set of circumstances. The scientist testified to the following facts (as given by Grey J in *R v Drummond* (2015)):

(1) The DNA samples from the clothing of the defendant excluded the complainant as a contributor.
(2) The DNA samples from the clothing of the complainant excluded the defendant as a contributor.

[1] Another explanation for the findings could be that a male relative of T's friend was the source of the second minor component and they would not have appeared as a separate male contributor on the Y-STR analysis. However, there was no suggestion that a male relative of T's friend was involved and so it has not been considered.

[2] DNA-IQ™ is a DNA extraction method that utilises the DNA adsorption properties of silica. Magnetic particles, coated in silica, are applied to cellular material that has been broken open to release the DNA. The DNA (now adsorbed onto the particles) is retained in the sample, while the undesired material is washed away. Heating releases the DNA from the particles, allowing it to be used in DNA profiling.

[3] Profiler Plus™ is a commercially available DNA profiling kit that targets 10 regions on the human genome, one of which is a sex-determining marker.

(3) The likelihood of DNA being left on a surface is dependent in part on the nature of the surface, the nature of the contact with that surface and a person's propensity to shed DNA.
(4) DNA is more likely to be left on a surface such as wood or fabric than a surface such as glass.
(5) DNA is more likely to be left on a surface where there has been prolonged or vigorous contact.
(6) A failure to obtain DNA from an item does not preclude the possibility that contact with that item occurred.
(7) The DNA testing conducted on the clothing of the complainant and the defendant did not preclude contact having taken place between the complainant and the defendant.
(8) A small study into the success of sampling at Forensic Science SA disclosed that DNA that may be uploaded[4] on to the database is recovered in about 10% of cases. This study related to samples where it was unknown whether DNA had in fact been left on each sample and, as a result, was only a 'sort of indication'[5] of how useful DNA samples are for uploading onto the database.

No likelihood ratios (*LR*) were provided by the scientist for the comparison of Drummond or T to the DNA profiles obtained from samples of their tops.

It was also brought up during the scientist's testimony that results may also be consistent with the prosecution version of events if Drummond had grabbed T by the arm and not contacted T's top.[6]

At the conclusion of the trial Drummond was found guilty and sentenced to 5 years and 3 months imprisonment.

1.4 *The source of the 10% figure*

It was the value of 10% in point 8 in the previous section and its understood meaning in the minds of the jury that has been the cause of the appeal and so it is worth explaining the source of the value. The figure mentioned by the forensic scientist came from the appendix of the report (included as a standard appendix in all reports).[7]

The source of the 10% figure given in the appendix is from a study (Sly and Sifis, 2008) carried out at FSSA in the no-suspect (database) section. This study examined 'success rates' (as defined by obtaining a profile that could be loaded to the database, termed 'uploadable' profiles) for different sample types falling under the umbrella of 'contact DNA' using the methodologies current to FSSA in

[4] Profiles may be submitted (termed 'uploaded') to a national list of profiles generated as part of criminal casework. Various criteria surround the suitability of a profile to be uploaded, but in the case of FSSA at the time the criterion was that a profile from a single contributor could be determined from at least 6 regions, out of the 10 being tested.

[5] As described by the scientist in her oral testimony.

[6] Note that this is an explanation of the results once that they have been obtained (post hoc explanation). Importantly, one cannot assign the value of the results given an explanation (see Evett *et al.*, (2000); More on the hierarchy of propositions: exploring the distinction between explanations and propositions. Science & Justice 40, 3 - 10). Indeed, one would then have to assess the value of results given explanations which are themselves based on the results (this would be a circular argument). Thus for the scientist to assess the value of the observations, propositions must be formulated before knowing the results of the comparison.

[7] The appendix stated:
Contact DNA – Contact DNA refers to biological material left on an object through a contact transfer (such as touching, handling or wearing) and not via deposition of a biological fluid (such as blood, semen or saliva). Contact DNA samples usually contain only small quantities of DNA and therefore analyses often do not give an informative profile. FSSA studies have found that of the majority of commonly submitted contact DNA sample types, only about 10% yield an informative DNA profile.

2008 (Chelex[8] based extraction (Walsh *et al.*, 1991) and using the Profiler Plus™ amplification kit). Later in 2012 this data collation was repeated (Nguyen *et al.*, 2012) for samples extracted using DNA-IQ (the same extraction methodology as used for the samples in the *R* v. *Drummond* matter).

The 10% figure given in the appendix of the report was obtained as the approximate value of the average of the success rate (obtaining an uploadable profile) of all samples in the 2008 study (actual value is 12%, but was rounded to 10% for the report appendix).

1.5 *The appeals*

The two FSSA studies, the transcript of the trial and the DNA reports of the scientist were sent to two independent experts hired by defence counsel. The following were raised by the experts:

(1) The more appropriate study to use was the 2012 study, which used the same extraction methodology as the samples in the *R* v. *Drummond* matter (it is worth noting here that at the time of the trial, in 2010, the 2012 study had not yet been conducted and so was unavailable to the testifying scientist)
(2) The more appropriate value is that for clothing specifically as clothing was the item in question. This would provide a success rate of 24% rather than 10%
(3) The case at hand did not have the primary purpose of obtaining a profile that could be loaded to a database, the definition of success for the case at hand would be any sample that yielded a profile with information that could be used for comparison to a reference. Referring to the 2012 FSSA study this corresponds to a value of 91%.
(4) Even then the values in the FSSA studies have no relevance because they are the product of casework samples and hence do not represent controlled experiments where the circumstances of deposition are known (a fact that the defence council brought up in cross examination of the forensic scientist during the trial).

The defendant lodged an appeal in 2012, which was dismissed. In 2013, an application for permission to appeal a second time, based on the information above, was made *R v. Drummond* (2013) and ultimately refused. The single judge determining the application rejected it for a number of reasons but in large part because the figure of 10% was appropriately explained by the forensic scientist and had only been used to support the scientist's ultimate proposition that it is not guaranteed that contact will result in a DNA being deposited. The defendant then appealed against the decision of the single judge.

In these appeals and applications the defence argument was that the 10% figure provided by the scientist was confusing and misleading to the jury and that they would have been left with the impression that there was only a 10% chance that Drummond's DNA would have been detected even if he had attacked T, whereas the true chance of detecting someone's DNA under such circumstances may be substantially different from than this.

In 2013 legislation changed in South Australia to allow second appeals based on fresh and compelling evidence (section 353A of the Criminal Law Consolidation Act 1935). This allowed the second appeal to be lodged. In this instance the decision of two of the three judges was that the evidence given by the forensic scientist was potentially misleading and the conviction should be overturned. In September 2015 the charges against Drummond were withdrawn and the court set aside his conviction.

[8] Chelex® is a DNA extraction method whereby a chelating agent is added to a sample to bind molecules that break down DNA. Once added, the solution is heated to break open cells and release DNA. The chelating agent and cellular material is then compressed into a pellet by centrifugation, allowing the DNA in solution to be removed and used in DNA profiling.

Now that the circumstances that led to the current state of affairs have been explained I will go through various concepts and issues that explain the difficulties encountered in this case.

2. Evaluating the findings

2.1 *The hierarchy of propositions*

A well-accepted and described concept in forensic evidence is the hierarchy of propositions (Cook *et al.*, 1998). This concept describes the various levels at which questions can be posed leading to the ultimate issue. The levels in the hierarchy are; offence, activity, source and sub-source. The scientist evaluates the results given the proposition useful to the court, and the court evaluates the propositions knowing the evidence. In Table 1 are examples of proposition pairs that can be considered in the *R* v.

TABLE 1 *Propositions and examples that could be considered in the* R *v.* Drummond *case*

Hierarchy	Notation	Proposition
Offence	$^{Offense}Hp$	Drummond attempted to kidnap T
	$^{Offense}Hd1$	No attempt was made by anyone to kidnap T
	$^{Offense}Hd2$	A male other that Drummond attempted to kidnap T*
Activity	$^{Activity}Hp$	T and Drummond struggled which included Drummond grabbing T's arm and T hitting Drummond's chest†
	$^{Activity}Hd1$	No-one grabbed or struggled with T
	$^{Activity}Hd2$	A male other than Drummond struggled with T which included them grabbing T's arm and T hitting them on the chest
Source‡	$^{Source}Hp1$	Drummond has contributed biological material (skin or sweat) to the top of T
	$^{Source}Hd1$	Drummond has not contributed biological material (skin or sweat) to the top of T
	$^{Source}Hp2$	T has contributed biological material (skin or sweat) to the top of Drummond
	$^{Source}Hd2$	T has not contributed biological material (skin or sweat) to the top of Drummond
Sub-source	$^{Sub\text{-}Source}Hp1$	Drummond has contributed DNA to the sample from the top of T††
	$^{Sub\text{-}Source}Hd1$	Drummond has not contributed DNA to the sample from the top of T
	$^{Sub\text{-}Source}Hp2$	T has contributed DNA to the sample from the top of Drummond
	$^{Sub\text{-}Source}Hd2$	T has not contributed DNA to the sample from the top of Drummond

*Although the main defence proposition during the trial aligns with Hd_2, I will explore the effects of both Hd_1 and Hd_2 in this article.

†While the activity as given in $^{Activity}Hp$ would be suspicious it does not in itself constitute a crime and hence is not an offence level proposition.

‡The propositions should be set by prosecution and defence and in this instance there was no indication that propositions relating to the source of biological material were being considered by either party. I therefore give the source level propositions as examples of propositions at this level, noting that they are not part of the consideration of the case.

††The propositions I have given are general with respect to the DNA profiling results to which they could be applied. In a typical evaluation of DNA results the analyst would choose more defined propositions, specific to each DNA profile obtained, for example the result from T's top may be evaluated with sub-source level propositions:

$^{Sub\text{-}Source}Hp1$: The sources of DNA are T, T's friend and Drummond

$^{Sub\text{-}Source}Hd1$: The sources of DNA are T, T's friend and an unknown male

Drummond case, specifically in relation to the clothing examined. Propositions are given in pairs that correspond to the prosecution stance (*Hp*) and the defence stance (*Hd*).

In this instance the scientist excluded T from the DNA obtained from the sample of Drummond's top and Drummond from the DNA obtained from the sample of T's top. Effectively the scientist has made the decision that the probability of obtaining these results if any of $^{Source}Hp1$, $^{Source}Hp2$, $^{Sub\text{-}Source}Hp1$ or $^{Sub\text{-}Source}Hp2$ is true, is zero and hence no *LR* was calculated, because it is therefore also set to zero. In this matter the sub-source or source levels are the positions in the hierarchy that the scientist is able to comment on, taking into account only the DNA profiling results. The question posed by prosecution and defence however relate to the probability of obtaining these results giving competing activities. In other words they were asking questions at the activity level, which required the scientist to consider more than just the DNA profiles obtained, and consider in addition the mere presence or absence and the quantity of detected DNA.

To help answer questions at the activity level the scientist requires information on the transfer of biological material, specifically:

(1) The probability that biological material will be transferred by grabbing clothing
(2) The probability that biological material will be transferred by hitting clothing

Before considering the points above it is worth first exploring the various definitions of success that have been suggested in this matter.

2.2 *What is success?*

Clearly the value of 10% as I have explained it in this article does not address any of the probabilities of interest regarding transfer, given in the previous section. This fact was pointed out by both defence exerts in their affidavits and was also explained by the forensic scientist during their testimony, however if in the minds of the jurors this 10% related somehow to the activity level propositions then it could constitute a potential miscarriage of justice.

As mentioned previously it was noted by defence experts the chance of transfer may be significantly higher than 10% (up to 91%). This figure still does not inform us of the value that is required to assess the findings with regard to activity level propositions. One of the major limitations with using the data from FSSA studies is that many of the DNA profiles obtained from the clothing samples in the 2012 study are likely to be those of the owners, and deposited over a long period of contact with the garment. The issues surrounding the use of data from either the 2008 or 2012 study to address the question of interest was noted by both defence experts, who stated that they '*. . . can have no relevance because they are the product of casework samples . . .* ' and hence are not carried out under controlled conditions where the true nature of the contact is known. The Judges in the second appeal also noted the lack of applicability of these findings when it was noted that '*The figures given for the testing of clothing relate to both DNA transferred to the clothing by the person wearing the clothing and DNA transferred to the clothing by a person other than the wearer of the clothing*'.

Now that we have identified which figures are not relevant to the question of interest to the court the question remains what information is required to help address the activity level propositions.

Data is required that relates to the percentage of short, one-off grabs of cloth that yield the DNA profile (or part thereof) of the grabber. We require the same sort of information relating to hitting rather than

grabbing. Such data suggests that experiments carried out under controlled conditions are required, and preferably using the same (or as close as possible) laboratory techniques as used in the case at hand.

2.3 *The probabilities of interest*

I turn now to the forensic literature in order to obtain the probabilities of interest in this matter. The studies I will rely on do not replicate in every detail the circumstances of the alleged activities in *R* v. *Drummond*, but do provide a solid foundation to assign the value of the evidence given competing activity level propositions. Before obtaining these values a formal setting for using them must be established and to do so requires some inevitable formula derivation.

We seek the ratio of the probability of the DNA evidence, E, given the competing propositions and the relevant information I (note that in all equations following this one I is omitted from the formula for ease of notation).

$$LR = \frac{\Pr\left(E|^{Activity}Hp, I\right)}{\Pr\left(E|^{Activity}Hd, I\right)}$$

I will decompose the evidence into parts, defining the results as:

$\overline{D_{D\to T}}$—There was an absence of Drummond's reference in the profile detected on T's top
$\overline{D_{T\to D}}$—There was an absence of T's reference in the profile detected on Drummond's top
$\overline{D_{U\to T}}$—There was an absence of the unknown male's reference in the profile detected on T's top (applicable when considering $^{Activity}Hd2$)
B_{T0}—The presence of background DNA on T's top
B_{D0}—The presence of background DNA on Drummond's top
I now consider two *LRs*, one for each of the two activity level proposition pairs:

$$\begin{aligned} LR_1 &= \frac{\Pr\left(\overline{D_{D\to T}}, \overline{D_{T\to D}}, B_{T0}, B_{D0}|^{Activity}Hp\right)}{\Pr\left(\overline{D_{D\to T}}, \overline{D_{T\to D}}, B_{T0}, B_{D0}|^{Activity}Hd\right)} \\ &= \frac{\Pr\left(\overline{D_{D\to T}}, \overline{D_{T\to D}}|B_{T0}, B_{D0}, {}^{Activity}Hp\right)\Pr\left(B_{T0}, B_{D0}|^{Activity}Hp\right)}{\Pr\left(\overline{D_{D\to T}}, \overline{D_{T\to D}}|B_{T0}, B_{D0}, {}^{Activity}Hd\right)\Pr\left(B_{T0}, B_{D0}|^{Activity}Hd\right)} \end{aligned}$$

$$\begin{aligned} LR_2 &= \frac{\Pr\left(\overline{D_{D\to T}}, \overline{D_{T\to D}}, B_{T0}, B_{D0}|^{Activity}Hp\right)}{\Pr\left(\overline{D_{D\to T}}, \overline{D_{T\to D}}, \overline{D_{U\to T}}, B_{T0}, B_{D0}|^{Activity}Hd_2\right)} \\ &= \frac{\Pr\left(\overline{D_{D\to T}}, \overline{D_{T\to D}}|B_{T0}, B_{D0}, {}^{Activity}Hp\right)\Pr\left(B_{T0}, B_{D0}|^{Activity}Hp\right)}{\Pr\left(\overline{D_{D\to T}}, \overline{D_{T\to D}}, \overline{D_{U\to T}}, |B_{T0}, B_{D0}, {}^{Activity}Hd_2\right)\Pr\left(B_{T0}, B_{D0}|^{Activity}Hd_2\right)} \end{aligned}$$

In *LR1* the defence proposition i.e. that T did not struggle with anyone, and had no contact with Drummond during or before the alleged incident. In LR_2 the defence proposition is that T struggled with an unknown male (not Drummond). Since the DNA on T's top is accounted for by T, T's male friend and an unknown female individual, it can be concluded that if an unknown male offender has grabbed T's arm then there is an absence of their DNA detected in the mixed profile from T's top ($\overline{D_{U\to T}}$). In these instances the probability of obtaining background DNA on the clothes of T and Drummond

are equal given either proposition, i.e. $\Pr(B_{T0}, B_{D0}|^{Activity}Hp) = \Pr(B_{T0}, B_{D0}|^{Activity}Hd) = \Pr(B_{T0}, B_{D0}|^{Activity}Hd_2)$. If we can also consider $\overline{D_{D\to T}}$, $\overline{D_{T\to D}}$ and $\overline{D_{U\to T}}$ to be independent of B_{T0} and B_{D0}, this gives:

$$LR_1 = \frac{\Pr(\overline{D_{D\to T}}, \overline{D_{T\to D}}|^{Activity}Hp)}{\Pr(\overline{D_{D\to T}}, \overline{D_{T\to D}}|^{Activity}Hd)}$$

$$LR_2 = \frac{\Pr(\overline{D_{D\to T}}, \overline{D_{T\to D}}|^{Activity}Hp)}{\Pr(\overline{D_{D\to T}}, \overline{D_{T\to D}}, \overline{D_{U\to T}}|^{Activity}Hd_2)}$$

We can also break the evidential findings apart so that:

$$LR_1 = \frac{\Pr(\overline{D_{D\to T}}|\overline{D_{T\to D}}, {}^{Activity}Hp)\Pr(\overline{D_{T\to D}}|^{Activity}Hp)}{\Pr(\overline{D_{D\to T}}|\overline{D_{T\to D}}, {}^{Activity}Hd)\Pr(\overline{D_{T\to D}}|^{Activity}Hd)}$$

$$LR_2 = \frac{\Pr(\overline{D_{D\to T}}|\overline{D_{T\to D}}, {}^{Activity}Hp)\Pr(\overline{D_{T\to D}}|^{Activity}Hp)}{\Pr(\overline{D_{D\to T}}|\overline{D_{T\to D}}, \overline{D_{U\to T}}, {}^{Activity}Hd_2)\Pr(\overline{D_{T\to D}}|\overline{D_{U\to T}}, {}^{Activity}Hd_2)\Pr(\overline{D_{U\to T}}|^{Activity}Hd_2)}$$

In words, the numerator of the likelihood ratio is equal to the probability of an absence of Drummond's reference in the profile from T's top if they struggled and given that there was an absence of T's reference in the profile from Drummond's top, multiplied by the probability of finding an absence of T's reference in the profile from Drummond's top (again if they struggled). The denominator is the same but considering the probability of the results in light of the defence proposition that T did not struggle with Drummond or anyone else (LR_1) or an unknown male which also resulted in an absence of their DNA profile from T's top (LR_2).

I make the further assumption that under the prosecution proposition the probability of an absence of Drummond's reference in the profile from T's top does not depend on whether there was an absence of T's reference in the profile from Drummond's top. Note that there are potentially high order dependencies that are not being taken into account when making this assumption, for example if there is an absence of T's reference in the profile from Drummond's top then this might suggest that the struggle was brief, which would then increase the chance of there being an absence of Drummond's reference in the profile from T's top. I ignore this potential dependency as (if it indeed exists) it is likely to be slight and doing so maximises the favour of the evidence for the defence (and hence could be described as a conservative assumption). Under the defence propositions $\overline{D_{D\to T}}$, $\overline{D_{T\to D}}$ and $\overline{D_{U\to T}}$ are independent because Drummond and T have not been in contact with each other.

Therefore:

$$LR_1 = \frac{\Pr(\overline{D_{D\to T}}|^{Activity}Hp)\Pr(\overline{D_{T\to D}}|^{Activity}Hp)}{\Pr(\overline{D_{D\to T}}|^{Activity}Hd)\Pr(\overline{D_{T\to D}}|^{Activity}Hd)}$$

$$LR_2 = \frac{\Pr(\overline{D_{D\to T}}|^{Activity}Hp)\Pr(\overline{D_{T\to D}}|^{Activity}Hp)}{\Pr(\overline{D_{D\to T}}|^{Activity}Hd_2)\Pr(\overline{D_{T\to D}}|^{Activity}Hd_2)\Pr(\overline{D_{U\to T}}|^{Activity}Hd_2)}$$

I now introduce terms for transfer:
$T_{D\to T}$—a transfer of biological material from Drummond to T.
$\overline{T_{D\to T}}$—no transfer of biological material from Drummond to T.
$T_{T\to D}$—a transfer of biological material from T to Drummond.
$\overline{T_{T\to D}}$—no transfer of biological material from T to Drummond.
$T_{U\to T}$—a transfer of biological material from an unknown male to T.
$\overline{T_{U\to T}}$—no transfer of biological material from unknown male to T.

Consider how the introduction of these transfer events affects the elements of the two *LR*s being developed. For example considering relevant transfer events to the first element of the numerator $\Pr\left(\overline{D_{D\to T}}|^{Activity}Hp\right)$ yields:

$$\Pr\left(\overline{D_{D\to T}}|\overline{T_{D\to T}},{}^{Activity}Hp\right)\Pr\left(\overline{T_{D\to T}}|^{Activity}Hp\right)+\Pr\left(\overline{D_{D\to T}}|T_{D\to T},{}^{Activity}Hp\right)\Pr\left(T_{D\to T}|^{Activity}Hp\right)$$

This formulation can be simplified by making the following assumptions:

- The probability of finding an absence of an individual's reference in the profile from an item if no DNA of the individual was transferred is one, i.e.$\Pr\left(\overline{D_{X\to Y}}|\overline{T_{X\to Y}},{}^{Activity}H\right)=1$[9].
- The probability of there being an absence of someone's reference from a sample given that material was transferred is zero, i.e. $\Pr\left(\overline{D_{X\to Y}}|T_{X\to Y},{}^{Activity}H\right)=0$[10].
- In other elements of the *LR*s we can make the simplifying assumption that the probability of no transfer of DNA occurring between Drummond and T given the defence proposition (of no contact) is 1, i.e. $\Pr\left(\overline{T_{D/T\to T/D}}|^{Activity}Hd\right)=1$ and therefore the probability of transfer is zero, $\Pr\left(T_{D/T\to T/D}|^{Activity}Hd\right)=0$.

Applying transfer events and the simplifications outlined above yields the *LRs*:

$$LR_1=\Pr\left(\overline{T_{D\to T}}|^{Activity}Hp\right)\Pr\left(\overline{T_{T\to D}}|^{Activity}Hp\right)$$

$$LR_2=\frac{\Pr\left(\overline{T_{D\to T}}|^{Activity}Hp\right)\Pr\left(\overline{T_{T\to D}}|^{Activity}Hp\right)}{\Pr\left(\overline{T_{U\to T}}|^{Activity}Hd_2\right)}$$

There is one additional simplification that can be considered when evaluating *LR2* and i.e. that the probability of a transfer (or no transfer) of material from someone, whether it is Drummond or an unknown male, given the scenario of a struggle is constant. Therefore $\Pr\left(\overline{T_{D\to T}}|^{Activity}Hp\right)=\Pr\left(\overline{T_{U\to T}}|^{Activity}Hd_2\right)$ and:

$$LR_2=\Pr\left(\overline{T_{T\to D}}|^{Activity}Hp\right)$$

[9] Note that this is a simplification, in reality the probability would be slightly less than 1 given that there is a chance of some of the background alleles matching but due to the very small effect this would have on the *LR* I use 1.

[10] This is also a simplification. In fact what we are assuming is that the probability of an absence of someone's reference from a DNA profile given that material was transferred, persisted on the item until sampling and recovered during laboratory analysis, is zero. Due to difficulty in experimentally distinguishing these events they are commonly combined under the single 'transfer' event, as is the case here. I consider the effect of persistence separately later.

We therefore need probabilities:

$\Pr(\overline{T_{D\to T}}|^{Activity}Hp)$—the probability of no transfer from Drummond to T in the manner described by prosecution

$\Pr(\overline{T_{T\to D}}|^{Activity}Hp)$—the probability of no transfer from T to Drummond in the manner described by prosecution.

For the probability of transfer to fabric I use the result of Daly *et al.* (2013) where cloth samples were held by volunteers for 60 s before sampling using tapelifts (the same sampling technique as using in the samples taken for *R* v. *Drummond*) and extracted using Qiagen® QIAamp DNA mini kit.[11] Table 1 of the Daly study shows that 53 out of 100 samples produced less than 0.01ng/μL of DNA and I will use this as an approximate cut-off for obtaining DNA profile information for Profiler Plus™.[12] There is no distinction between holding and hitting, therefore I will assume $\Pr(\overline{T_{D\to T}}|^{Activity}Hp) = \Pr(\overline{T_{T\to D}}|^{Activity}Hp) = 0.53$. Another study into the rate of DNA transfer to clothing in a simulated struggle was carried out by Sethi *et al.* (Sethi *et al.*, 2013). In this study 'assailant' volunteers were asked to grab the arms, elbows and wrists of 'victim' volunteers for 15 s, while they struggled to free themselves. Three types of cloth were grabbed (cotton, polyester and a cotton/polyester blend) and sampled after 12 h and 7 days. DNA extraction was carried out on 0.5cm^2 swatches of cut cloth using QIAamp DNA micro kit. From the 27 samples tested at 12 h only three yielded a detectable amount of DNA.[13] This is approximately 11% of samples. I will use a transfer rate of 0.53 from the first study for two reasons; The sample size in the Daly *et al.* (2013) study is much larger than the Sethi *et al.* (2013) study and the sampling technique is equivalent to that used in the *R* v. *Drummond* matter (i.e. tapelifts).[14]

Applying a probability of non-transfer results in:

$$LR_1 = \Pr(\overline{T_{D\to T}}|^{Activity}Hp)\Pr(\overline{T_{T\to D}}|^{Activity}Hp) = (0.53)^2 = 0.28$$

$$LR_2 = \Pr(\overline{T_{T\to D}}|^{Activity}Hp) = 0.53$$

2.4 *Consideration of DNA persistence*

Note that I have not considered persistence of DNA in my evaluation. There is little data available on the persistence of trace DNA. There are a number of factors that are likely to affect persistence, such as

[11] This is an extraction methodology whereby cells are broken open and DNA is adsorbed onto a silica membrane. Impurities are washed through the membrane and the DNA is then eluted into a solution to be used in DNA profiling.

[12] The Daly paper also goes on to produce DNA profiling results, but uses a different DNA profiling system to that in the *R* v. *Drummond* matter and hence those results are less applicable than the quantification result. Only fabric samples that yielded 0.03 ng/μL were profiled in the Daly study due to the low probability of obtaining a useable (defined by the Daly *et al.* authors as obtaining 6 alleles they could attribute to a single contributor) DNA profile. The quantification value if 0.01ng/μL was used as a cut-off in this manuscript because a) it was the lowest quantification category recorded in the Daly study, and b) it is the author's experience with using Profiler Plus™ that DNA profiles would regularly not be obtained from less that this concentration of DNA.

[13] Note that from a contact such as that in the Sethi study we would expect there to always be some transfer of DNA, however in many instances the amount of DNA that has transferred and persisted until sampling will be below the limit of detection of the system. Hence I refer to a lack of 'detectable' DNA rather than a lack of DNA.

[14] Note that I am not promoting the finding of the Daly *et al.* (2013) study as generic result for all considerations of DNA transfer from hand to cloth. It may be that case-specific circumstances will require the findings from different work, or that a laboratory may need to carry out their own controlled experiments to better match the details of the alleged crime. In this case the larger the probability of a DNA transfer, the more support the *LR* will ultimately provide to the defence proposition, and so the findings from the Daly *et al.* (2013) study will yield an *LR* that provides more support to the defence proposition than the Sethi *et al.* (2013) study. I discuss these aspects further in the 'The reliability of activity level reporting' section.

the surface type, the length of time and the conditions the item is exposed to during the time. The best example of a DNA persistence study for contact DNA is the work by Raymond *et al.* (2009). In this study known amounts of biological material were deposited on items and either kept outdoors or indoors over the course of approximately 40 days. The amount of DNA recovered was examined at daily intervals. While there is a high degree of variability in the amount of DNA that was detected, an examination of the findings in their study show that very low (or even no detectable) levels of degradation occur within the first few days (the timeframe that is relevant in the Drummond matter). Given the circumstances of the case and the timeframe of sampling persistence is unlikely to have a major influence on the size of the *LR*. Nevertheless it is a worthwhile exercise to investigate how DNA persistence could be incorporated into the *LR*, and its potential effect.

I start with the formulae for LR_1 and LR_2, prior to the implementation of the probabilities of transfer. Retaining the previously justified assumptions:

- $\Pr\left(\overline{D_{X\to Y}}|, \overline{T_{X\to Y}}, {}^{Activity}H\right) = 1$
- $\Pr\left(\overline{T_{D/T\to T/D}}|{}^{Activity}Hd\right) = 1$
- $\Pr\left(T_{D/T\to T/D}|{}^{Activity}Hd\right) = 0$

and simplifications:

- $\Pr\left(T_{D\to T}|{}^{Activity}Hp\right) = \Pr\left(T_{T\to D}|{}^{Activity}Hp\right) = \Pr\left(T|{}^{Activity}Hp\right)$
- $\Pr\left(\overline{D_{D\to T}}|{}^{Activity}Hp\right) = \Pr\left(\overline{D_{T\to D}}|{}^{Activity}Hp\right) = \Pr\left(\overline{D}|{}^{Activity}Hp\right)$

yields:

$$LR_1 = \left[\Pr\left(\overline{D}|, T, {}^{Activity}Hp\right)\Pr\left(T|{}^{Activity}Hp\right)+\Pr\left(\overline{T}|{}^{Activity}Hp\right)\right]^2$$

$$LR_2 = \Pr\left(\overline{D}|, T, {}^{Activity}Hp\right)\Pr\left(T|{}^{Activity}Hp\right)+\Pr\left(\overline{T}|{}^{Activity}Hp\right)$$

I now consider persistence probabilities:

P—the DNA persisted on the clothing of Drummond and T in detectable levels.
$\overline{P}$—the DNA did not persist on the clothing of Drummond and T in detectable levels.[15]
Incorporating these into the *LR* gives:

$$LR_1 = \left[\begin{array}{l}\Pr\left(\overline{D}|, T, P, {}^{Activity}Hp\right)\Pr\left(T|P, {}^{Activity}Hp\right)\Pr(P|Hp)+\\ \Pr\left(\overline{D}|, T, \overline{P}, {}^{Activity}Hp\right)\Pr\left(T|\overline{P}, {}^{Activity}Hp\right)\Pr\left(\overline{P}|Hp\right)+\Pr\left(\overline{T}|{}^{Activity}Hp\right)\end{array}\right]^2$$

$$LR_2 = \Pr\left(\overline{D}|, T, P, {}^{Activity}Hp\right)\Pr\left(T|P, {}^{Activity}Hp\right)\Pr(P|Hp)+$$
$$\Pr\left(\overline{D}|, T, \overline{P}, {}^{Activity}Hp\right)\Pr\left(T|\overline{P}, {}^{Activity}Hp\right)\Pr\left(\overline{P}|Hp\right)+\Pr\left(\overline{T}|{}^{Activity}Hp\right)$$

Making the new simplifications that:

- The probability of finding an absence of an individual's reference in the profile from an item if DNA was transferred and persisted is zero, i.e. $\Pr\left(\overline{D}|, T, P, {}^{Activity}H\right) = 0$

[15] Again, I do not distinguish between persistence of Drummond's DNA on T's top and T's DNA on Drummond's top. In some circumstances, for example where a marked difference in garment properties, such a distinction would be warranted.

- The probability of finding an absence of an individual's reference in the profile from an item if DNA was transferred and didn't persist is one, i.e. $\Pr(\overline{D}|, T, \overline{P}, {}^{Activity}H) = 1$
- The probability of a transfer occurring is independent of the probability that any transferred material would persist i.e. $\Pr(T|\overline{P}, {}^{Activity}H) = \Pr(T|{}^{Activity}H)$

gives:

$$LR_1 = \left[\Pr(T|{}^{Activity}Hp)\Pr(\overline{P}|Hp)+\Pr(\overline{T}|{}^{Activity}Hp)\right]^2$$

$$LR_2 = \Pr(T|{}^{Activity}Hp)\Pr(\overline{P}|Hp)+\Pr(\overline{T}|{}^{Activity}Hp)$$

In words, the group of probabilistic terms that make up LR_1 and LR_2 can be thought of as considering the absence of an individual's DNA from an object given the prosecution proposition as either by the fact that DNA was transferred but did not persist on the object or that the DNA was not transferred to the object.

While I have derived these formulae using explanations in terms of DNA, they are typical of well documented formula that deal with these same issues of transfer and persistence in non-DNA evidence such as fibres (Champod and Taroni, 1992) or glass (Curran *et al.*, 2000). In fact the formulae for LR_2 yields that same formula from Hicks *et al.* (2016) by a simple nomenclature definition, $\Pr(\overline{T_{T\rightarrow D}}|{}^{Activity}Hp) = t_0$. A similar incorporation of the transfer and background probabilities considering blood stains at crime scenes or on individuals is shown in Aitken and Taroni (2004).

As mentioned previously, there is little information in the literature regarding the persistence of contact DNA on clothing that can inform on a value to use for $\Pr(P)$. In the absence of such information the effect of a range of values for $\Pr(P)$ (i.e. from 0 to 1) on the *LRs* can be considered, while holding $\Pr(\overline{T}|{}^{Activity}Hp) = 0.53$ constant (Fig. 1).

The vertical axis in Fig. 1 has been given as the *LR*. To represent the result as a level of support for the defence activity level proposition compared to the prosecution activity level proposition the *LR* values need to be inverted. As the probability that DNA persists becomes lower the value of the evidence is driven neutrality. As it becomes more probable that DNA would persist, if it were

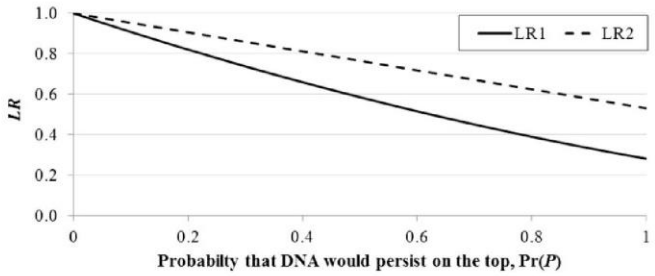


FIG. 1. LR over the range of persistence probability values.

transferred, the support for the defence proposition over the prosecution proposition increases. Conservatively the value for $\Pr(P)$, or $\Pr(\overline{P})$, in this matter could be chosen to maximize support for the defence proposition, which would be $\Pr(P) = 1$.

What has been carried out here is known as a sensitivity analysis with respect to DNA persistence. In a sensitivity analysis the assigned value of probabilities are varied across a range of sensible values they could take in order to assess the effect on the size of the *LR*. In this way sensitivity analyses allow us to understand the limitations in the calculation due to the paucity of data in a given study. In the analysis of evidence in R v Drummond I have now considered the effect that incorporating a probability of persistence in the calculation has on the resulting *LR*.

Such a sensitivity analyses can be carried through to consider the effect of probability assignments for more than factor on the *LR*. For example, if I wished to investigate the effects of both persistence and transfer on the size of the *LR* then I would obtain graphs (now in three dimensions) as shown in Fig. 2.

The graphs shown in Fig. 2 for LR_1 and LR_2 are very similar, and the keenly observant may notice a slightly more gentle dip of values for LR_1 travelling towards the bottom right of the graphs. As the probability of transfer and persistence both approach certainty together the support for the defence proposition increases. At the extreme if transfer and persistence of DNA were certain then this would also signify a certainty that the prosecution scenario could not have occurred. However, this dramatic support for the defence proposition only occurs at the most extreme values of transfer and persistence; for most values, even relatively high compared to literature findings, the support for the defence propositions remains slight. Also note that the probability of an absence of observable transferred DNA can never be higher given the prosecution proposition than given the defence proposition. The absence of evidence in this case is indeed evidence for absence, contradictory to what the old adage may suggest.

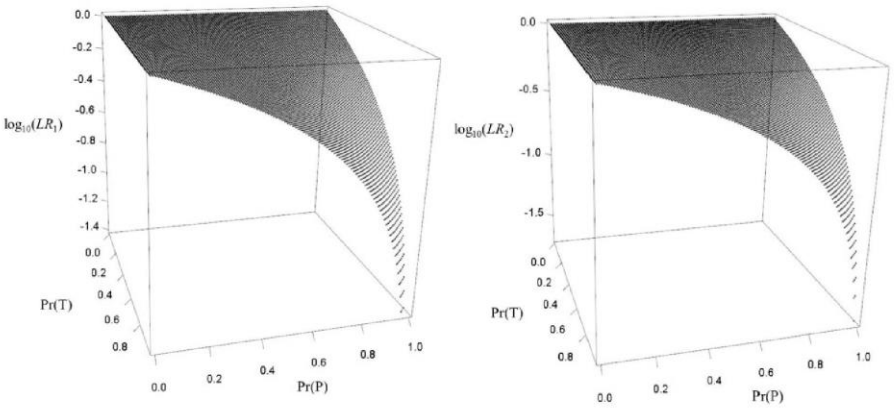


FIG. 2. LR values, shown in log10 scale when considering a range of assigned probabilities for transfer and persistence.

2.5 *The strength of the DNA findings*

When asked the significance of the DNA results in the *R* v. *Drummond* matter it would be possible to provide a statement considering either set of activity level propositions. I provide an example statement below that takes both proposition pairs into account.

> There is an absence of Drummond's reference in the profile obtained from the tapelift of the top of T. There is an absence of T's reference in the profile obtained from the tapelift of the top of Drummond. In the evaluation of the evidence I have considered two possible propositions for these results;
>
> - The prosecution proposition I have considered is that T and Drummond struggled which included Drummond grabbing T's arm and T hitting Drummond's chest.
> - The defence proposition I have considered is that T had no prior contact, direct or indirect, with Drummond (and either did not struggle with anyone, or struggled with an unknown male).
>
> When I consider the probability that DNA would be transferred in such an encounter, the DNA findings are in the order of 2 to 4 times more likely to have been obtained if the defence proposition[16] had occurred rather than prosecution proposition. This provides slight (or weak or limited) support to the proposition that T had no prior contact, direct or indirect, with Drummond (and either did not struggle with anyone, or struggled with an unknown male), compared to the proposition that T and Drummond struggled.
>
> The propositions were formed from the information available to me at the time. If this information changes or if the defence or prosecution nominate alternative propositions then I will need to re-valuate the findings.

3. Discussion

3.1 *The reliability of activity level reporting*[17]

On reading the calculations in this work the reader may be sceptical of the validity of the assumptions that have been made and the applicability of the values obtained from scientific literature sources. A common observation of activity level reporting is that there are too many variables to consider in the evaluation of any particular case and that no controlled experiment could ever hope to recreate them closely enough to be applicable. From the outset this appears to be a strong argument and indeed, as can be seen from the calculations in this article, a number of simplifying assumptions need to be made in order to make the problem tractable. There will be cases where the complexity or ambiguity surrounding the alleged activities, or the general lack of relevant published data or personal knowledge means that an activity level assessment should not be attempted. However when this is the case, it may not be appropriate to fall back to source or sub-source level propositions. In a report by the European Network of Forensic Institutes (Willis, 2015) on evaluative reporting they highlight this point by stating '*Source level propositions are adequate in cases where there is no risk that the court will*

[16] Either $^{Activity}Hd_1$ or $^{Activity}Hd_2$.

[17] The term 'activity level reporting' can imply that reports are made on the probabilities of activities having occurred. This is not the case. We report the probability of findings given activity level propositions, we do not report on the probabilities of activities themselves.

misinterpret them in the context of the alleged activities in the case.'. As might be gleaned from this statement (and from the fact that internationally agreed guidelines exist for activity level reporting) the practise of activity level reporting is considered highly desirable within the forensic community. There are two further points to be made here.

Firstly, it is the author's experience that the results of DNA profiling or their interpretation are not often challenged in court, increasingly so with the introduction of modern DNA profile interpretation systems that have the ability to analyse highly complex DNA profiles and take much of the subjectivity out of DNA profile interpretation. Instead, the interest of the court focuses on how the presence of an individual's DNA can be explained on an item through suggested activities (either innocent or incriminating). If the scientist carries out an activity level assessment, as outlined in this article, they will be fully aware of the assumptions being made, the limitations with the findings, the available literature on the topic and the logical framework in which the problem sits. The scientist is in the best position to help guide the court in assessing the value of the evidence given various posited activities. If challenged on the source of their opinion a clear trail of reasoning is available for scrutiny. Alternatively an activity level assessment might not be done on the grounds that there is too much complexity to consider. Whether or not such an assessment has been carried out the questions in court are still going to be asked, leaving the scientist only able to answer questions with statements such as '*that is possible*' or '*in my opinion that is unlikely*', but without specifically being able to show the supporting evidence of that opinion. This has been raised previously by Champod (2013). Note that by giving an answer such as '*in my opinion that is unlikely*' the scientist has provided a comment on the probability of the activity level propositions, but without elucidating a structured scientific reasoning for the opinion with a transparent foundation. The scientist who considers activities in the evaluation of all their evidence is one who will be better prepared to answer questions in court.

The second point to consider is that no amount of controlled experimental work will be able to replicate the exact circumstances surrounding an alleged activity. For most alleged activities the exact circumstances are simply not known (or indeed may not ever have occurred). Consider a simple example relevant to the case at hand; it may be that controlled experiments to assign the probability of transfer from hand to cloth during a struggle will simulate struggles that are too vigorous or apply too much pressure compared to the alleged event and so overestimate the probabilities that biological material will be transferred. Conversely it may also be that they are not vigorous enough compared to the alleged event and underestimate the transfer probabilities. Either way the results obtained from the controlled experiments (matched as closely as possible to what is known about the alleged activities) will represent the best information available for the probabilities pertinent to the case. For any numerical assessment of the evidence these results are required, and the only alternative, if they are not available, is to provide an experience based estimate[18] of what the scientist believes is the probabilistic value. Again, having some data that can guide our beliefs on transfer is better than having none.

Notwithstanding the last two points, there are certain criteria that a controlled experiment should meet. The controlled experiment being carried out (or the study from which the value is being taken) needs to have some level of similarity with the alleged activities. There is no value in using data from studies that are completely removed from the circumstances of the case, and at worst could result in the scientist providing misleading evidence to the court. It is up to the scientist to make clear in their report or testimony the assumptions they have made and any differences between the studies used as models

[18] This is when the scientist provides their best estimate for a probability without relying on any specific literature. Such experience based estimates are sometimes referred to as 'soft data'.

for the alleged activity and the alleged activity itself. If this information is made clear then it is open for the court to challenge these assumptions and studies and the effect they could have on the *LR*. This leads to another aspect of assigning an *LR*, which I touched on in consideration of persistence of this work, called sensitivity analyses.

It must be remembered that the court will use this evidence to update its prior beliefs of innocence or guilt.[19] The *LR* I have calculated in the *R* v. *Drummond* matter provides slight support to the proposition that T had no prior contact, direct or indirect, with Drummond (and T either did not struggle with anyone, or struggled with an unknown male), compared to the proposition that T and Drummond struggled and the court's decision would likely be dominated by its prior beliefs accumulated through other evidence. Nevertheless, the presentation of relatively weak or neutral evidence remains an important finding to present. Even if providing completely neutral evidence the scientist has at least informed the court that they must rely entirely on other evidence to support their beliefs of guilt or innocence.

3.2 *Can we comment on more than activities?*

There are a number of considerations in the *R* v. *Drummond* matter which relate to issues that sit separately from the disputed activities, such as:

- The relevance of the sample from T's top to the alleged offence (i.e. whether it was grabbed)
- The number of offenders (i.e. whether there were zero or one offender)
- The possibility for innocent transfer to have occurred between Drummond and T prior to the offence

Consideration of these facets of the evidence is generally thought to require offence level propositions. Evett (1993) introduced terms *k*, *a* and *r* for the number of offenders, the possibility of innocent transfer and the relevance of exhibits, respectively, and demonstrated their incorporation into a *LR* equation that considers propositions:

- The suspect was one of the k offenders
- The suspect was not one of the k offenders

Whether the inclusion of these considerations by the scientist in their evaluation of the findings helps the court or usurps its role is a matter of some debate. In his work Evett (1993) explains that '*Unless the circumstances are very precisely defined, it may be that the appropriate probabilities for r and a must be the province of the court*', and I will expand on this shortly.

A similar set of considerations was explained in Aitken *et al.* (2003) who considered terms *t*, *p*, *r* and *b* for:

- Transfer, persistence and recovery
- Innocent acquisition
- Relevance and
- Innocent presence (commonly referred to as background).

[19] While this is unlikely to happen in a numerical sense, it is the general form of Bayes' rule within which scientific evidence is presented.

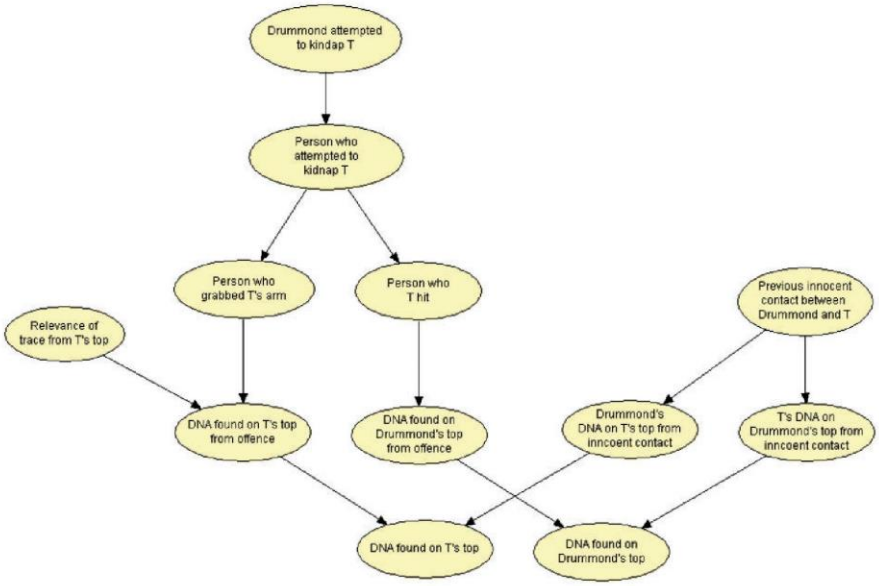


FIG. 3. Bayesian network considering the scenario at hand.

Incorporation of multiple findings and case circumstance information can be complex, particularly when doing so purely by formulaic derivation. A tool exists that displays these considerations graphically, with arcs showing dependencies between the factors. This tool is called a Bayesian Network (*BN*) and I direct the reader to the works of Taroni et al. (2014) for an excellent exploration of their construction and use. *BN* have been used to assess findings in light of propositions at offence (Evett *et al.*, 2002), activity (Biedermann *et al.*, 2009; Gittelson *et al.*, 2012; Wieten *et al.*, 2015) and more recently source (Taylor *et al.*, 2016, Wolff *et al.*, 2015) and sub-source (Biedermann *et al.*, 2012; Dawid *et al.*, 2006; Mortera, 2002) levels.

In Fig. 3 I show a *BN* that is designed to evaluate the findings in *R* v. *Drummond* given offence level propositions. For reasons of brevity I do not describe the construction of the *BN* in great detail, nor do I go into the population of the conditional probabilities that underlie each node.[20] For the interested reader I can provide more information on request. Again for the sake of brevity I ask the reader to accept that I have populated the nodes of the *BN* with appropriate probabilities regarding transfer and persistence (which in this *BN* are combined into a single conditional probability), so that I can continue directly on with the effects of considering relevance, innocent transfer and the number of offenders in the evaluation of the findings.

I start with a brief description of the *BN* in Fig. 3. There are two 'branches' on the *BN*, which ultimately meet in the nodes that relate to the results of the presence (or absence) of DNA. One branch

[20] A 'node' is shown graphically as an oval in Figure 3, and represents a random variable.

considers the possibility that DNA was transferred through a struggle and the other branch considers the means by which DNA could have come to be transferred between Drummond and T for innocent means; in this case a previous contact. The node that aligns with the offence level propositions is the parent of the struggle branch, and has an arc directly to the '*Person who attempted to kidnap T*' node. In this node I take into account that it could have been Drummond, and unknown male or no-one that attempted to kidnap T (i.e. whether there were one or no offenders). This node then leads to the nodes that relate to the expected transfers of DNA to the tops of Drummond and T. The final node of note is one which considers the relevance of the trace taken from T's top. In particular the relevance of the sample depends on the probability that the alleged offender contacted T's top at all, or just her arm.

In the Bayesian Network (*BN*) construction I make the following assumptions:

- The absence of DNA from Drummond on T's top and DNA from T on Drummond's top is not disputed.
- The probability of an absence of Drummond's DNA from T's top does not depend on whether there was an absence of T's DNA from Drummond's top. The two are considered independent events. Note that this is different from what is commonly referred to as cross-transfer (Aitken *et al.*, 2003).[21]
- The background DNA on the top of T and Drummond are not relevant to the offence. Note that this is the reason for the lack of nodes for the innocent presence of DNA (or background DNA).
- I have not considered the possibility of laboratory error in this evaluation. It is a common practise to make the conservative assumption of a zero false negative rate with such an analysis (which is what has been done here). Had some inclusionary findings been obtained there has been a number of publications demonstrating the incorporation of the possibility of a false positive error, e.g. see (Thompson *et al.*, 2003).

There are three probabilities that underlie the nodes in the *BN* for which there could be a question of whether they impinge on the role of the court, although possibly two of these could readily be accepted as uncontroversial. The first of these is the probability of a previous contact between Drummond and T. While in some trials, such information could be disputed and rely on eyewitness statements or alibis, in the Drummond matter it was accepted by both parties that there had been no previous contact, either direct or indirect between Drummond and T. The second probability to discuss is the relevance of the trace from T's top. It is difficult to assign a prior probability to the relevance of this trace and in the absence of any other information (such as damage on the top, eye-witness testimony, or any level of certainty on the part of T) I have assigned equal prior probabilities for the alleged offender either grabbing T's arm or her top. This probability assignment is also likely to be uncontroversial, as neither party would seem to have strong view as to state of reality.

The final probability I wish to mention, and probably the most controversial, is whether any attempt to kidnap T was made (i.e. whether there were zero or one offender). This probability is used in the '*Person who attempted to kidnap T*' node and in the *BN* in Fig. 3 I have again used equal prior probabilities for these states. These probability assignments are likely to be based on the submissions put forward by prosecution and defence, the testimony of those involved and the believability of their

[21] It is different as the two potential transfers result from two different activities. In fact the dependence between the two activities has been considered due to the joint parental node '*Person who attempted to kidnap T*'. Instantiation of one of the activity nodes ('*Person who grabbed T's arm*' or '*Person who T hit*') sees the other activity node receive a probability of 1 on the corresponding state. An example of cross transfer in the classic sense would be if, when considering the alleged hitting of Drummond by T, samples from both Drummond's top and T's hand had been taken and examined.

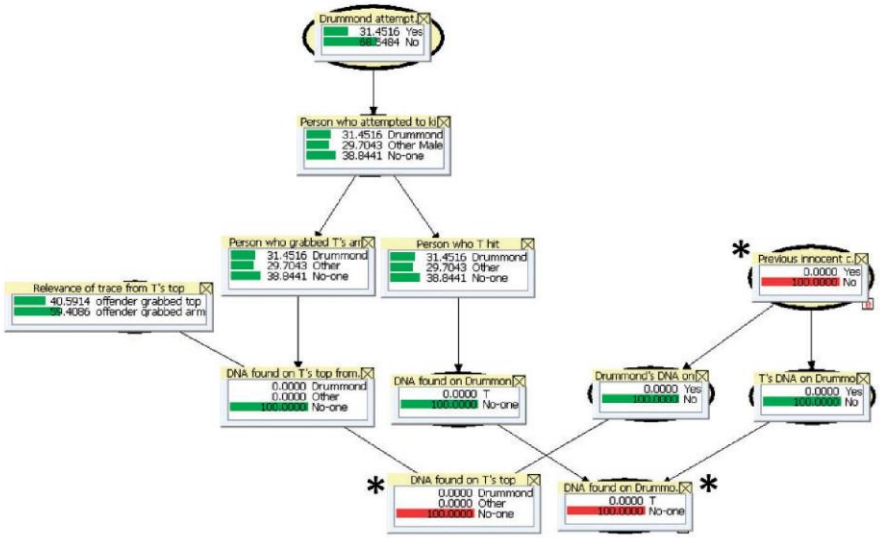


FIG. 4. BN seen in Fig. 3 with case specific information instantiated (marked with an asterisk).

testimony. As such it could well be argued that these probabilities are the province of the court and should not be assigned by the scientist. Even attempts to lead jurors through a process of logical Bayesian inference, using numerical values for similar considerations, have been the cause for appeals in the past, see *R. v D.J. Adams (1997)*. Nevertheless, for the purposes of demonstrating the use of *BN* to help address the findings given offence level propositions I continue, arbitrarily using equal prior probabilities.

Having populated the tables for each node within the *BN* the following information can be provided:

- No DNA from Drummond was present on the top of T.
- No DNA from T was present on the top of Drummond.
- There was no previous contact between Drummond and T.

Figure 4 shows the *BN* from Fig. 3, with this information provided (in *BN* parlance referred to as instantiation). The instantiated information is shown in nodes marked with an asterisk in Figure 4, and from the propagation of this information through the nodes in the *BN* the resulting posterior probabilities for each state in the remaining nodes can be seen in green bars.

Note that the probabilities shown in the '*Drummond attempted to kidnap T*' node in Fig. 4 are posterior probabilities as they are a combination of the prior information provided to the states within this node and the information provided at other nodes propagated through the *BN*. In the *BN* I have constructed, the two states that the '*Drummond attempted to kidnap T*' node can take are given equal prior probabilities. The *LR* that is produced by the division of probabilities from Fig. 4 is $\frac{\Pr(Hp|E)}{\Pr(Hd|E)}$, however as the priors are equal, $\Pr(Hp) = \Pr(Hd)$, the *LR* is equal to the posterior probability ratio, i.e.

$\frac{\Pr(Hp|E)}{\Pr(Hd|E)} = \frac{\Pr(E|Hp)}{\Pr(E|Hd)}$. In this matter, $\frac{\Pr(Hp|E)}{\Pr(Hd|E)} = \frac{\Pr(E|Hp)}{\Pr(E|Hd)} = \frac{31.4516}{68.5484} \approx 0.4588$. Or if the propositions are inverted, $\frac{\Pr(E|Hd)}{\Pr(E|Hp)} \approx 2.18$.

The conclusion that could then be provided is:

Given the propositions that either:

- Drummond attempted to kidnap T.
- Drummond did not attempt to kidnap T.

Where it is claimed that during the alleged offence the right arm or shoulder of T was grabbed by the offender and T hit the offender's chest. I have considered that it may be the case that no attempt was made to kidnap T. I have evaluated the probability of finding no DNA of T on Drummond's top and no DNA from Drummond on T's top. The ratio of the probability of these findings giving each of the two competing propositions has been assigned as approximately 0.5.

Given the information available to me, my view is that the findings are in the order of two times more probable given that Drummond did not attempt to kidnap T as opposed to attempting to kidnap her.

It can be seen that the *LR* assigned when considering offence level propositions in this matter is not markedly different than that obtained when considering activity level propositions.

3.3 *Carrying out research to help address disputed activities*

I end this section with some thoughts about ongoing work in the area of activity level reporting. In general the fewer assumptions that need to be made in the evaluation of the evidence, the closer it will mirror the alleged activities (but the more complex it will become). One possibility when helping to evaluate the findings considering activity level propositions would be to design and carry out experiments that as closely match the circumstances of each alleged activity. In reality, time and resource pressures will make such a policy unworkable for most practising forensic laboratories and so controlled experiments must be designed that can be applied to multiple cases. I present some ideas for doing so:

- If results are reported based on the interpretation of DNA profiles e.g. high level mixtures, dominant contributions, more than an arbitrary minimum number of alleles present, etc., they will rely on the specific functioning of the profiling kit used in the study and hence cannot be applied to other situations that use different profiling kits. Instead, the results could be reported as an amount of DNA of the person of interest present in a sample. This finding can still be calculated by using information the DNA profile result (i.e. the relative height of peaks that correspond to that person), but related back to the initial amount of DNA detected on the sample. Modern DNA profile analysis software can provide estimates on the proportion of individuals within a mixture which facilitates this approach. The report of a DNA amount is independent of the DNA profiling kit or process used to generate the profile and so can be applied to a broader range of calculations.
- Due to the difficulties inherent in separating the aspects of transfer, persistence and recovery most controlled experiments base their findings on the combined effects of all three. This makes each experiment very specific to the case at hand and difficult to apply to any other. Where possible these effects should be separated. For example, the scientist may be interested in the probability of obtaining a DNA profile from blood as a result of a fist fight and after the garment has been

washed. Rather than carrying out an experiment whereby people fist fight, wash their clothes and then submit them for profiling the experiment could be broken up into three parts; a) the probability and amount of blood transferred during fist fighting, b) the persistence of blood through washing and c) the recovery of DNA from blood stains on garments. Some of these components could then be applied in other scenarios, such as a stabbing where blood stained clothing was washed. While this seems like an obvious example there are instances (particularly when dealing with trace DNA) where separating transfer, persistence and recovery is difficult. Breaking experiments into modules in this manner means that complex case scenarios can be constructed from series of smaller experiments, minimizing the required additional work in each scenario.

- When possible carry out regression analyses on the data that relate the measured outcome (e.g. such as DNA amount, mixture status, level of degradation, etc) to the variables tested in the study (e.g. such as amount of starting DNA, length of exposure to some environmental condition, length of contact, vigour of contact, etc). Firstly, it may be that one or more of the variables are found to have no bearing on the final result. This finding provides important information upon which scientists can base their assumptions. Alternatively if dependence is found then a regression analysis may allow extrapolation of the data to combinations of values of the existing variables that were not included as part of the original study (keeping in mind that extrapolation will be informative only so far, as the further out from the observed range the extrapolation extends the wider the distribution of possible values will be in order to accommodate the lack of information). Access to the results of regression analyses also allows linking of findings of separate studies, i.e. you may be able to combine the results of a study on DNA transfer from varying lengths of contact with a result on the DNA transfer from varying vigorousness of contact by their regression results.

4. Conclusion

The testimony, the affidavits and the appeal judgements in the *R* v. *Drummond* case highlight important concepts when evaluating DNA profiling results, particularly exclusions. The case demonstrates the importance of recognizing different levels in the hierarchy of propositions and the information required to evaluate a *LR* at these levels. It also highlights the importance of choosing appropriate statistics when supporting an argument as, even if properly explained, the chance that they are misconstrued can lead to appeal.

Appropriate statistical treatment of the data shows the exclusionary DNA evidence can be evaluated in light of activity level propositions. While the presence of someone's DNA (or support for its presence) is generally accepted in court within the framework of a case, the significance of the absence of DNA is more difficult to comprehend. The results in this case show that the findings provide slight support for the defence propositions ($LR_1 = 0.28$ and $LR_2 = 0.53$) compared to the prosecution account. Sensitivity analyses showed that for reasonable levels of persistence and transfer the strength of the evidence in favour of the defence proposition remained slight. Even when taking offence level considerations of relevance, number of offenders and innocent means of DNA transfer into account the strength of the evidence still remained slight and in favour of defence.

Acknowledgements

Points of view in this document are those of the author and do not necessarily represent the official position or policies of Forensic Science SA. I gratefully acknowledge guidance and helpful discussions with Tacha Hicks Champod and Christophe Champod, whose contribution greatly improved this work. I also thank two anonymous reviewers and editor whose comments were invaluable.

References

1997. R. v. D.J. Adams (EWCA crim 222).

2013. R v Drummond. Supreme Court of South Australia - SASCFC 135.

2015. R v Drummond (No. 2). Supreme Court of South Australia - SASCFC 82.

Aitken, C. G. G. and Taroni, F. 2004. *Statistics and the evaluation of evidence for forensic scientists*, 2nd edition. John Wiley & Sons, Ltd.

Aitken, C. G. G., Taroni, F. and Garbolino, P. 2003. A graphical model for the evaluation of cross-transfer evidence in DNA profiles. *Theoretical Population Biology* **63**, 179–190.

Biedermann, A., Bozza, S., Konis, K. and Taroni, F. 2012. Inference about the number of contributors to a DNA mixture: Comparative analyses of a Bayesian network approach and the maximum allele count method. *Forensic Science International: Genetics* **6**, 689–696.

Biedermann, A., Bozza, S. and Taroni, F. 2009. Probabilistic evidential assessment of gunshot residue particle evidence (Part I): Likelihood ratio calculation and case pre-assessment using Bayesian networks. *Forensic Science International* **191**, 24–35.

Champod, C. 2013. DNA transfer: informed judgement or mere guesswork? *Frontiers in Genetics* **4**, 1–3

Champod, C. and Taroni, F. 1992. Interpretation of Fibre Evidence. In J. Robertson and M. Grieve (eds), Taylor & Francis: London.

Cook, R., Evett, I. W., Jackson, G., Jones, P. J. and Lambert, J. A. 1998 A hierarchy of propositions: Deciding which level to address in casework. *Science and Justice* **38**, 231–240.

Curran, J. M., Hicks, T. N. and Buckleton, J. 2000. *Forensic Interpretation of Glass Evidence*. CRC Press: London.

Daly, D. J., Murphy, C. and McDermott, S. D. 2013. The transfer of touch DNA from hands to glass, fabric and wood. *Forensic Science International: Genetics* **6**, 41–46.

Dawid, A. P., Mortera, L. and Vicard, P. 2006. Object-oriented Bayesian networks for complex forensic DNA profiling problems. *Forensic Science International* **169**, 195–205.

Evett, I. W. 1993. Establishing the evidential value of a small quantity of material found at a crime scene. *Journal of the Forensic Science Society* **33**, 83–86.

Evett, I. W., Gill, P. D., Jackson, G., Whitaker, J. and Champod, C. 2002. Interpreting small quantities of DNA: the hierarchy of propositions and the use of Bayesian networks. *Journal of Forensic Sciences* **47**, 520–530.

Evett, I. W., Jackson, G. and Lambert, J. A. 2000. More on the hierarchy of propositions: exploring the distinction between explanations and propositions. *Science & Justice* **40**, 3–10.

Gittelson, S., Biedermann, A., Bozza, S., Taroni, F. 2012. Bayesian networks and the value of the evidence for the forensic two-trace transfer problem. *Journal of Forensic Sciences.* **57**, 1199–1216.

Hicks, T., Buckleton, J., Bright, J.-A. and Taylor, D. 2016. A Framework for Interpreting Evidence. In J. Buckleton , J.-A. Bright and D. Taylor (eds), CRC Press: Boca Raton, Florida.

Mortera, J. 2002. Analysis of DNA mixtures using Bayesian networks. In P. Green , N. L. Hjort and S. Richardson (eds), Oxford University Press: Oxford.

Nguyen, K., Sly, N., Henry, J. and Sifis, M. 2012. Success rates for trace and non-trace DNA samples extracted using Promega DNA-IQ. *21st International ANZFSS Symposium*. Hobart, Australia

RAYMOND, J. J., OORCHOT, R. A. H. V., GUNN, P. R., WALSH, S. J., and ROUX, C. 2009. Trace evidence characteristics of DNA: A preliminary investigation of the persistence of DNA at crime scenes. *Forensic Science International: Genetics* **4**, 26–33.

SETHI, V., PANACEK, E., GREEN, W.M., JILLIAN, N.G. and KANTHASWAMY, S. 2013. Yield of Male Touch DNA from Fabrics in an Assault Model. *Journal of Forensic Research*, T1: 001. doi:10.4172/2157-7145.T1-001.

SLY, N. and SIFIS, M. 2008. South Australian Casework Contact/Trace DNA Success Rates. *19th International ANZFSS Symposium*. Melbourne, Australia

TARONI, F., BIEDERMANN, A., BOZZA, S., GARBOLINO, P. and AITKEN, C. 2014. *Bayesian networks and probabilistic inference in forensic science*: 2nd edition. John Wiley & Sons, Ltd.: Chichester.

TAYLOR, D., ABARNO, D., CHAMPOD, C. and HICKS, T. 2016. Evaluating forensic biology results given source level propositions. *Forensic Science International: Genetics* **21**, 54–67.

THOMPSON, W. C., TARONI, F. and AITKEN, C. G. G. 2003. How the probability of a false positive affects the value of DNA evidence. *Journal of Forensic Science* **48**, 1–8.

WALSH, P. S., METZGER, D. A. and HIGUCHI, R. 1991. Chelex® 100 as a medium for simple extraction of DNA for PCR-based typing from forensic material. *BioTechniques* **10**, 506–513.

WIETEN, R., ZOETE, J. D., BLANKERS, B. and KOKSHOORN, B. 2015. The interpretation of traces found on adhesive tapes. *Law, Probability and Risk* **14**, 305–322.

WILLIS, S. 2015. ENSFI Guideline for Evaluative Reporting in Forensic Science. European Network of Forensic Science Institutes.

WOLFF, T. R. D., KAL, A. J., BERGER, C. E. H. and KOKSHOORN, B. 2015. A probabilistic approach to body fluid typing interpretation: an exploratory study on forensic saliva testing. *Law, Probability and Risk* **14**, 323–339.

Chapter 9: Impact of the work described in this thesis

The significance of the work outlined in this thesis can be noted in two ways. The first of these is the addition to the forensic biology statistics field in general, and particularly in the area of probabilistic genotyping. Many of the works that have been presented in this thesis have become the standard papers to cite when discussing aspects of probability-based DNA profile interpretation. Evidence of this is that the paper in section 2.6 'The interpretation of single source and mixed DNA profiles' has been cited 123 times and the paper in section 2.2 'Developing allelic and stutter peak height models for a continuous method of DNA interpretation' has been cited 89 times (as of February 2019). The publication of these methods and models came quite early in the sub-field of continuous probabilistic genotype systems and as such pioneered many of the ways that these systems should be developed, evaluated and implemented. Evidence of this is the heavy referencing of these articles in probabilistic genotyping validation and use guidelines published by the American advisory body SWGDAM (Scientific Working Group on DNA Analysis Methods) and the European based group ISFG (International Society of Forensic Genetics). The method of summarising and interpreting data has become a standard method for probabilistic genotyping software assessment, for example the scatter plots shown in the paper in section 5 'Using continuous DNA interpretation methods to revisit likelihood ratio behaviour' is now routinely used by numerous other groups in their assessments. My involvement in the development of the methods and models associated with the published works is outlined prior to each paper presented in this thesis. This work has most often been carried out in collaborations with colleagues all around the world.

Perhaps the best way to show a tangible impact of the work presented in the publications of this thesis is through their implementation in probabilistic genotyping software STRmix™. The manner in which STRmix™ came to be developed, and my involvement in that process is outlined in the various chapters of this thesis. To summarise; initially the mathematics and modelling that was developed for DNA profile analysis (as presented in this thesis, particularly chapter 1) was programmed by me (using the java programming language) into the software STRmix™. I, John Buckleton and Jo-Ann Bright were then involved in providing training to other laboratories (initially Australian, but then later overseas) on the use of this new method of profile evaluation. The paradigm shift in the way that profiles were analysed, combined with user feedback saw us refine and develop models for STRmix™, which were subsequently published and implemented into code (again by myself). As STRmix™ has grown over the years the need for dedicated (and professional) programmers became a requirement, and also the need for dedicated support staff. As o 2019 approximately 20 people are employed by the STRmix™ company and we retain the services of professional programmers full-time for development. I am still involved in programming, but this is now mainly to implement the science and mathematics additions, while leaving aspects of programming that deal with the interface, file manipulations, licensing and auditability to the professionals.

STRmix™ was introduced into active forensic biology casework in 2012 in Forensic Science SA and ESR (the two laboratories of the co-developers of the software). Since then the software

has become the standard to use on all Australian and New Zealand forensic laboratories and is currently (as of February 2019) in approximately 60 forensic laboratories, spanning the United States of America, Europe, Middle East and China. Figure 9.1 shows the rate of uptake of the STRmix™ software over the past 7 years.

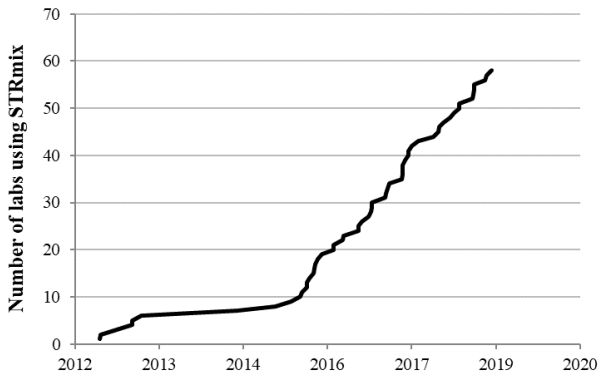


Fig 9.1: Graph showing number of laboratories using STRmix™

The biggest uptake of STRmix™ has been in the forensic biology laboratories in the USA, where now over half of the ANAB (standing for ANSI National Accreditation Board, where ANSI stands for American National Standards Institute) are using STRmix™. This includes the three federal laboratories associated with the FBI, the US Army and the Bureau of Alcohol, Tobacco, Firearms and Explosives, who have all since carried out and published their own research using STRmix™.

STRmix™ is now used routinely in forensic biology and university research projects and cited as a standard methods of DNA profile analysis by laboratories all around the world. I have personally been invited to laboratories around Australia, Dublin, UK, Northern Ireland, America and New Zealand and have been invited to present at workshops on this topic in Poland, Japan, New Zealand and around Australia as part of forensic and legal conferences.

In the last four years the use of STRmix™ has reached a level of maturity in the USA that (independent of the developers) a yearly STRmix™ conference is held (see Fig 9.2) to discuss aspects of the software, both theoretical and practical.

Additionally, large biotechnology companies host Webinar presentations on STRmix™ analysis and presentation of evidence in court.

It is impossible to gauge the number of casework samples that STRmix™ has been used to analyse worldwide. At FSSA, since its use was introduced in 2012, the number of analysed samples is in the order of 20 000. Given that FSSA is a relatively small laboratory (by forensic biology laboratory

standards) one could imagine that the worldwide figure would be in the hundreds of thousands.

Figure 9.2: flyer advertising the 2019 STRmix workshop

These profile analyses have made up the contents of numerous reports, provided to Police, Lawyers, Coroners, Private Investigators and Courts. The number of cases that have been influenced by the results of profiles analysed in STRmix™ is unknowable. I have been involved in testimony in Courts at the Magistrates, District and Supreme level for matters as minor as vandalism and as major as cold-case Homicides. One well-known case of note (for which I testified on the mathematics and modelling in STRmix™) is that of the murder of Louise Belle in 1983 in South Australia, for which STRmix™ was used and played a part in the conviction of Dieter Pfennig, over 30 years later in 2016.

Another case worth mentioning is that of Clinton Tuite, who was convicted in 2018 of a sexual assault in 2007. This case is particularly worth mentioning due to the level of challenge against many aspects of the evidence evaluation, including many aspects surrounding the development, mathematics, validation, use and validity of STRmix™. This matter spanned over four years in court starting in 2014 and due to the relatively new application of probabilistic genotyping at the time, it involved numerous voire dires testing the admissibility of different aspects of evidence. At its height the appeals reached the Court of Appeal, and the resulted in amendments to the Australian laws of forensic expert evidence to account for the new type of evidence being admitted into courts [2015 VSCA 148 - CLINTON TUITE V THE QUEEN].

As well as the significant contribution to court cases the models and methods within STRmix™ have been used for investigations, with the aim of identifying potential offenders. There are two aspects in particular where this has been shown. The first relates to a process known as ‘mixture searching’ (outlined in the article in section 3.4 ‘Searching mixed DNA profiles directly against profile databases’). This process of interrogating state or national databases for potential contributors to complex mixtures has been implemented in laboratories to varying degrees, from implementing the mathematics directly into the IT systems (as in ESR) or, mor commonly, using the searching function in STRmix™. At FSSA, the capability to search mixed DNA profiles over the years has (as of February 2019) resulted in investigative links being sent to SA Police for over 100 cases (with many instances of multiple links in each case) that would have otherwise never been possible to provide.

Figure 9.3: Promotion of a news article from the Advertiser 'Solving the impossible' February 8, 2019 written by Miles Kemp. Imagine is of the convicted offender Patrick Perkins

There is talk of going back through no-suspect cases examined prior to the implementation of mixture-searching in a large scale back-capture program, for which untold amounts of investigative information would be generated. The popularity of the mixture searching tool has lead to the development of separate software that is specifically designed with automated searching and auditing capabilities, so that they can be carried out in en masse.

In a similar theme of investigative searching, the mathematics outlines in the paper 'Considering relatives when assessing the evidential strength of mixed DNA profiles' in section 3.4 has been implemented to allow familial searching to be conducted. The use of the familial search function in STRmix™ lead to Australia's first conviction of an offender identified through such a search in South Australia (for the case of a serial stranger rapist). The result was widely publicised in the local media (Figure 9.3 shows a promo for a news article in the Advertiser).

The final point to note is the impact that the use of the mathematics and modelling within STRmix™ has impacted the general community. The level to which this has occurred is difficult to gauge. There will have been immediate impacts to those directly related to criminal investigations where STRmix™ has been used to analyse the DNA profile evidence. To the people more broadly the impact is felt through the feeling of safer community through better justice methods. The best way to demonstrate this is perhaps through the fact that there have been numerous articles in various public areas of newspaper, television, public events, or newsletters that speak to the improvement in DNA profile evaluation. Below I provide a selection of newspaper headlines that directly relate to STRmix™, and the mathematics being used therein, for the betterment of the community:

- MI Authorities Employ New DNA Analysis Software
- How DNA turns criminal's own family against them
- STRmix™ Use Leads to Indiana Murder Conviction
- New software can do what no human could, helps state police analyze DNA evidence
- Montreal forensics lab approves STRmix™ use

- Crime-Busting Forensic Software STRmix™ Triumphs in U.S. Murder Trial
- Houston-based Forensic Lab Approves Mixed DNA Profile Forensic Software
- New Mexico Case Allows Expert Testimony, Affirms STRmix™ Reliability
- DNA technology used to link convicted killer to another murder victim wasn’t available in 2009
- How new DNA technology led to Jupiter triple homicide arrests
- Eight More Agencies, Including ATF, Will Use STRmix™
- FL Murder Case Reaffirms Reliability of STRmix™
- Lost shoe led to landmark DNA ruling - and now, nation's 1st guilty verdict
- DNA breakthrough: North Adelaide rape suspect arrested
- ESR technology being used to solve war crimes
- 'Dream' software boost power of DNA