# Comparison of methods for forensic DNA typing of soils

**Anastasia S. Khodakova**
**B.Sc., M.Sc.**

**Thesis submitted for the degree of Doctor of Philosophy**

**February 2015**

**School of Biological Sciences**

**Flinders University**

**Adelaide, Australia**

# Table of contents

# Abstract


## Comparison of methods for forensic DNA typing of soils

Soil is encountered commonly during the course of a forensic examination yet rarely is it analysed. Currently, forensic comparison of soil evidence samples is limited to the analyses of mineral and elemental composition. Recent advances in high-throughput sequencing (HTS) opens the possibility for the most quantitative and accurate examination of genetic richness and diversity of soils. These methodologies provide the ability to generate unique metagenomic DNA profiles from a variety of soil types that can be applied to forensic soil discrimination.

The focus of this study is to compare the ability of modern metagenomic approaches to examine genetic similarities and variations of soil biota and to discriminate soils taken from different geographical locations a few km apart. The three most commonly used methods to study the whole biota include: shotgun sequencing where random DNA fragments of the whole metagenome are analysed; whole genome amplification (WGA) which is performed for sequencing of limited amounts of available DNA material; and arbitrarily-primed PCR (AP-PCR) where a single primer selectively amplifies sections within an entire metagenome prior to sequencing. In addition, gene-specific sequencing based on the evaluation of 16S rRNA bacterial genes was carried out.

Soil samples were taken from three different locations within the Adelaide urban residential area approximately 5 km apart. Two soils from similar land use and vegetation type could not be easily distinguished visually, while the third soil was of a distinctively different type. Replicate samples were collected to determine any variation during the year and reproducibility of data for samples collected from the same location.

Initial studies determined the optimum extraction and storage processes for retrieval of the maximum amount of high purity DNA from each sample type. The extracted DNA was analysed by each of the four metagenomic approaches using the Ion Torrent sequencing platform. The sequencing data was then analysed using traditional comparisons with different reference databases. Data from random whole metagenomic sequencing was additionally analysed using reference independent comparative bioinformatics approaches.

Multivariate statistical comparison of the soils' metagenomic profiles, obtained by AP-PCR and 16S rRNA-based sequencing techniques, allowed for accurate discrimination of the soil samples according to their geographical locations. Shotgun and WGA sequencing approaches generated highly similar profiles such that the soil samples could not be distinguished. The AP-PCR-based approach was shown to be successful at generating reproducible site-specific DNA profiles for samples collected from the same location through the different seasons of the year.

The results of our proof-of-concept study demonstrate for the first time that metagenomic PCR-based sequencing approaches are able to reliably discriminate between visually similar soil types sampled from close but different locations. This represents a significant step towards implementation of a metagenomic sequencing technique to discriminate soil samples for forensic practice.

# Declaration

I certify that this thesis does not incorporate without acknowledgement 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.

Anastasia S. Khodakova

February 2015

# Acknowledgments

It is my pleasure to thank many people who supported me during in my Ph.D. journey and made the completion of this thesis possible.

I would firstly like to thank my primary supervisor Professor Adrian Linacre for his mentoring, advices, assistance and support throughout this project. I am deeply grateful to you for giving me this opportunity to work in your research group and to pursue my scientific career. I appreciate all your contribution of time, good ideas, many fruitful discussions, encouragement and funding to ensure the project is completed. Your enthusiasm and versatile knowledge in forensic science and research were motivational for me.

I would also like to thank my co-supervisor Professor Leigh Burgoyne for his genuine interest in my research and sharing his vast knowledge on DNA typing methodologies and soil microbiology with me. I have learnt much from you.

I am thankful to Assoc. Prof. Damien Abarno for his involvement throughout the entire duration of this research project, valuable discussions on the research topic and for providing expert comments towards the writing and improvement of this thesis.

I owe a special thanks to Dr. Renee Smith for her experienced guidance in the world of metagenomics and statistical analysis of metagenomic data.

I would also like to thank Assoc. Prof. Duncan Taylor for his help towards my understanding of forensic statistics.

Many many thanks go to Forensic DNA group members: Sherryn Ciavaglia, Renee Blackie, Alicia Haines, Jennifer Templeton, Dr. Shanan Tobe, for your advices, support and friendship from the very beginning of my Ph.D.

I will never forget many other colleges and collaborators for their fruitful cooperation and assistance: Dr. Sophie Leterme, Dr. Paul Gooding, Assoc. Prof. Kathryn Burdon, Dr. Roman Dronov. I would like to express my sincere thanks to Professor Amanda Ellis for her constant support during my time at Flinders.

Though we have recently been out of touch, I am deeply grateful to Professor Natalia Beloborodova, who has awakened my interest in human microbiome research many years ago and taught me what good science and research should be. Her advices both in science and in life in general helped me to build up my confidence and encouraged me to set up my goals high.

Finally, I want to thank my parents who have supported me in every way throughout my life and career, who have been a constant source of wisdom and inspiration for me.

My biggest thank you of all goes to my husband Dr. Dmitriy Khodakov and our two adorable children. I am forever grateful for your love, patience, emotional support and care during all these years. Thank you for listening and being understanding, for believing in my strength and encouragement to move towards my goals. Without you I would not have come this far.

This thesis is dedicated to my family

# List of scientific achievements

## Publications

1. **Khodakova A.S.**, Burgoyne L.A., Abarno D., Linacre L. *Random Whole Metagenomic Sequencing for Forensic Discrimination of Soils.* **PLOS ONE, 2014**, 9(8): e104996. doi:10.1371/journal.pone.0104996
2. **Khodakova A.S.**, Burgoyne L.A., Abarno D., Linacre L. *Forensic analysis of soils using single arbitrarily primed amplification and high throughput sequencing*. **Journal of Forensic Science Internetional: Genetics Supplement Series, 2013,** doi:10.1016/j.fsigss.2013.10.019

**Additional publications (Appendix D)**

3. D.A. Khodakov, **A.S. Khodakova**, D. Huang, A. Linacre, A.V. Ellis. *Protected DNA strand displacement for enhanced single nucleotide discrimination in double-stranded DNA*. **Scientific Reports, 2015**, Accepted, In Press.
4. Khodakov D.A., **Khodakova A.S.**, Linacre A. and Ellis A.V. *Sequence selective capture, release and analysis of DNA using a magnetic microbead-assisted toehold-mediated DNA strand displacement reaction*. **Analyst**, 2014, 139(14): 3548-3551.
5. Khodakov D.A., **Khodakova A.S.**, Linacre A. and Ellis A.V. *Toehold-Mediated Nonenzymatic DNA Strand Displacement as a Platform for DNA Genotyping*. **JACS, 2013**, 135(15), 5612-5619.
6. Khodakov D.A., **Khodakova A.S.**, Linacre A. and Ellis A.V. *Amelogenin Locus Typing Using Toehold Assisted DNA Melting Curve Analysis*. **Journal of Forensic Science Internetional: Genetics Supplement Series, 2013**, doi:10.1016/j.fsigss.2013.10.061

# Oral Presentations

1. **25th World Congress of the International Society for Forensic Genetics (ISFG)** Melbourne, September 2013. *Application of next generation sequencing for forensic soil DNA analysis.*
2. **25th World Congress of the International Society for Forensic Genetics (ISFG)** Melbourne, September 2013. *Programmable DNA hybridisation as a new tool for forensic DNA genotyping.*
3. **22nd International Symposium on the Forensic Sciences (ANZFSS),** Adelaide, August 2014. *Metagenomic sequencing as a tool for forensic soil discrimination.*

# Poster presentations

1. **21st International Symposium on the Forensic Sciences (ANZFSS),** Hobart, September 2012. *Comparative analysis of commercial DNA extraction kits for the isolation of whole metagenome DNA from forensically relevant soil samples.*

# Abbreviations

| | |
|---|---|
| AAS | Atomic Absorption Spectroscopy |
| AGRF | Australian Genome Research Facility |
| ANOSIM | Analysis of Similarities |
| AP-PCR | Arbitrarily Primed PCR |
| ARDRA | Amplified Ribosomal DNA Restriction Analysis |
| ATR-FTIR | Attenuated Total Reflectance Fourier Transform Infrared Spectroscopy |
| BLAST | Basic Local Alignment Search Tool |
| BLAT | BLAST-like Alignment Tool |
| bp | Basepair |
| CAFSS | Centre for Australian Forensic Soil Science |
| CE | Capillary Electrophoresis |
| CLUSTER | Hierrahical Agglomerative Clustering |
| DAF | DNA Amplification Fingerprinting |
| DGGE | Denaturing Gradient Gel Electrophoresis |
| DNA | Deoxyribonucleic Acid |
| dNTP | DeoxyriboNucleotide TriPhosphate |
| EBC | Extraction Blank Controls |
| EDS | Energy Dispersive Spectroscopy |
| FAM | Fluoresceine |
| FBI | Federal Bureau of Investigation |
| FTIR | Fourier Transformed Infrared Spectroscopy |
| Gbp | Gigabase pair |
| HTS | High Throughput DNA Sequencing |
| ICP | Inductively Coupled Plasma Spectrometry |
| ICP-MS | Inductively Coupled Plasma-Mass Spectrometry |
| ICP-OES | Inductively Coupled Plasma-Optical Emission Spectrometry |
| IMG/M | Integrated Microbial Genomes |
| IPA | 2-propanol |
| ITS | Internal Transcribed Spacer |
| kbp | Kilobase pair |
| LH-PCR | Length-Heterogeneity PCR |
| LR | Likelihood Ratio |
| LSU | Large Subunit of Ribosomal RNA operon |
| Mbp | Megabase pair |
| MG-RAST | Metagenomic Rapid Annotations using Subsystems Technology |
| min | minute |
| mM | Millimolar |
| NAA | Neutron Activation Analysis |
| NCBI | National Center for Biotechnology Information |

| | |
|---|---|
| NMDS | Non-metric Multidimentional Scaling |
| OTU | Operational Taxonomic Units |
| PCR | Polymerase Chain Reaction |
| pdf | probability density functions |
| QIIME | Quantitative Insights Into Microbial Ecology |
| RAPD | Randomly Amplified Polymorphic DNA |
| RFU | Relative Fluorescent Units |
| RNA | Ribonucleic Acid |
| rRNA | Ribosomal Ribonucleic Acid |
| RT | Room Temp |
| s | second |
| SEM | Scanning Electron Microscopy |
| SIMPER | Similarity Percentages |
| SOP | Standard Operating Protocol |
| SSU | Small Subunit of Ribosomal RNA operon |
| STAMP | Statistical Analysis of Metagenomic Profiles |
| TGGE | Temperature Gradient Gel Electrophoresis |
| TRFLP | Terminal Restriction Fragment Length Polymorphism |
| UV-VIS | Ultraviolet–Visible Spectroscopy |
| WGA | Whole Genome Amplification |
| XRD | X-ray Diffraction |
| XRF | X-ray Fluorescence |
| μL | Microliter |
| μM | Micromolar |

# Chapter 1. General Introduction

## Soil analysis in forensic science

One of the most important tasks of any forensic investigation is to generate scientific support whether a suspect was, or was not, present at a crime scene. Through direct or indirect contact, people and objects deposit various types of trace evidence around a crime scene. The identification and examination of such traces include for example human hairs, blood, saliva, fingerprints, fibres from clothing - all of which can help to establish an evidential link that connects people or objects to a crime scene (Houck 2004). Many crimes occur in outdoor, urban and rural areas making trace evidence including geological, soil, botanic and other environmental materials potential source of forensic information (Ruffell & McKinley 2005; D. Pirrie et al. 2013). For example, particles of soil adhering to a suspect's car tyres may later be shown to be related to the soil from the crime scene thus supporting the assertion that the suspect's vehicle was present at the crime scene. Equally, soil traces found under a suspect's fingernails may be consistent with the soil specimen from a victim's dress, thereby establishing that some contact between the victim and suspect had occurred. As such, soil materials may be used for intelligence purposes, assisting a criminal inquiry or for comparative purposes that ultimately can lead to presentation of soil as evidence in court (Ritz et al. 2009).

## 1.1 History of the use of soil materials as trace evidence

The potential value of soil material in criminal investigations was recognised more than a century ago (Bergslien, 2012). Since then analysis of sediment and soil particles as a type of trace evidence has been investigated widely and has evolved to a high level of scientific sophistication and quality. However, due to the lack of a validated method, or a set of techniques, there is no acceptance of soil science in today's criminal justice system and it is rarely presented in court trials.

Perhaps the very first application of soil as an evidential material was demonstrated by one of the earliest forensic scientists, German chemist, Georg Popp in 1904 (Ray Murray, 2011). He examined the composition of soils found on suspects' clothing, shoes and in fingernail scrappings and then compared it with the composition of soil from the crime scene; this aided in the investigation of a murder case (Murray 2011; Bergslien 2012).

In France, Edmond Locard, a director of the Lyon Police Technical Laboratory, also developed an interest in the analysis of dust traces. In 1920, his work led to the development of what became known as the Locard's Exchange Principle, familiar to all who work in forensic science. The principle, postulating that "every contact leaves a trace", soon became established as the basis for modern approaches to the examination of all trace evidences, from fingerprints and fibres to soils (Murray 2011).

In the U.S. the use of soil material as trace evidence was introduced as early as the 1920s by the criminologist Edward Heinrich, who applied his expertise in geology to the investigation of crime cases. In the following ten years a set of approaches for soil and mineral analysis, proposed by Heinrich, gained wide acceptance by the Federal Bureau of Investigation (FBI) (Heinrich 1965). Other important forensic work was focused on the investigation of minerals and other particles found in soil using ultra-microanalysis techniques. Walter McCrone developed microscopical and microchemical techniques to study a wide range of particle types, which resulted in publication of the first Particle Atlas (McCrone et al. 1967).

In the UK scientists at the Home Office Laboratory at Aldermaston performed a remarkable series of studies in forensic geology between the 1950s and 1980s. These studies tested and showed the value of a number of physical and chemical techniques of soil analysis including microscopic examination, pH determination, colour description, sieving and heavy mineral analysis amongst other materials. At the same time, the lack of reliability of some methods such as the density gradient method was also demonstrated (Murray 2011).

In 1975, the first book on forensic geology was published by Murray and Tedrow where the authors described the advances made in forensic geoscience and illustrated them based on numerous case studies (Murray & Tedrow 1975). Since then progress in the use of forensic geology in crime scene investigation has been supported by many publications in peer-reviewed journals and police reports; for review see Pirrie *et al.* (Rawlins et al. 2006; Pye 2007; D Pirrie et al. 2013).

The exponential growth in the area of forensic human DNA analysis in the mid-1990s decreased the demand of forensic investigators in traditional trace evidences including soil. This in turn resulted in a decrease in the number of experienced soil forensic examiners as well as validated examination procedures (Pye 2007). During the past decade interest in the soil examination has increased further (Zala 2007). It is important to note that the majority of soil forensic investigations in Australia are performed by soil scientists from academia and not at state-funded forensic services (Woods et al. 2014).

In 2003 the forensic geologist Rob Fitzpatrick founded the Centre for Australian Forensic Soil Science (CAFSS) (http://www.clw.csiro.au/cafss/). The Centre has assisted with more than 100 crime cases in Australia and overseas

(Fitzpatrick & Raven 2012). Fundamental research conducted by CAFSS and the experience gained from these case studies allowed the development of the "Guidelines for conducting criminal and environmental soil forensic investigations" (Fitzpatrick & Raven 2013).

Moreover, the fact that soil analysis has gained a great deal of interest for modern criminal investigations is supported by the introduction of international soil forensic conferences and similar meetings where leading world experts discuss recent developments, progress and scientific community understanding of new and existing geo-analytical methods for soil analysis in forensic science.

Nevertheless, examination of soil as an intelligence tool for producing evidential material for court trials has limited implementation in routine forensic practice. Though, with the changing investigative environment the value of soil examination in parallel with other forms of physical intelligence is gaining increased recognition.

## 1.2 The nature of soil forensic samples and their limitations

Despite the ubiquitous nature of soil which is found at many crime scenes, it is rarely accepted as valuable forensic evidence. A reason of this could be the incorrect interpretation of soil evidence given soil is easily transferred over large distances. These multiple soil movements could mislead forensic investigators resulting in an incorrect reconstruction of the crime scene.

Forensic soil science and traditional earth sciences investigate the same objects and gain the same primary information. However, the philosophical approaches underlying the two differ fundamentally (Morgan et al. 2006). In order to

provide accurate forensic interpretation from the analysis of soil evidence it is important to recognise this difference between conventional science and forensic procedures.

A large number of soil samples encountered by a forensic investigator come from locations such as municipal parks, areas around the home, beaches, parking lots and rural properties. Soils of these types have a low value for general academic research because of high contamination and frequent changes caused by everyday human activities. In forensic science the term ‘soil’ means earth surface material including not only natural constituents but also artificial or exogenous components e.g. fibres, plastics, paints, metals, glass, bricks, *etc.* Significant variations in such soils make soil sampling a very challenging task for a forensic investigator.

Forensic soil analysis revolves around the comparison of soil samples in order to determine their provenance. Samples submitted for forensic investigation are categorised in different ways. The sample associated with a suspect or victim attracts the most attention and is referred as a questioned sample. Samples collected from a crime site are called control samples whose origin is assumed to be known. Alibi samples, whose origin is also known, are often collected from alternative locations that the suspect reported visiting. The fourth group of samples constitutes reference samples – the samples typically held in a museum or soil/geological archive collection (Murray 2004; Fitzpatrick & Raven 2012; Pye 2007).

### 1.2.1 Soil sampling

One of the major aspects that make soil examination for forensic purposes different to that performed for traditional soil science is the relatively small size of questioned samples. Examination of a trace soil sample (of the order of few

milligrams) that may not accurately represent the source material from which it is derived, is a challenging task facing the forensic scientist. One of the investigative tasks during an investigation where soil traces are recovered is to collect control samples from known locations that appear visually as close as possible to the questioned soil sample. The expectation being that at least one of those known samples will be consistent with the questioned sample. Random sampling is often the best way of obtaining an unbiased set of representative soil samples from the crime scene and control sites, however in some cases it may be reasonable to employ purposeful targeted sampling. The number of samples that should be taken from the control site depends on the size of the site and the specific nature of the environment at and around the site (Pye 2007).

It is anticipated that soil sampling has to be done in such a way to ensure representativeness of all soil samples involved in the investigation (Gilbert & Pulsipher 2005). Comparison of a limited size questioned sample with a large amount of sample from the control sites is far from trivial as an investigator must be aware of issues regarding comparison of trace samples to bulk samples that can contain soils from different horizons (Morgan & Bull 2007).

### 1.2.2 Soil transfer and mixing

Soil samples collected from the belongings and clothing of the victim or suspect are highly likely to contain materials from a number of different sources associated with pre- and post-crime contacts (French et al. 2012). Primary transfer occurs when, for example, an item directly makes contact with a particular source of evidence during a criminal act (Figure 1.1). Secondary transfer may occur if this item makes a secondary contact and transfers evidence obtained from the primary transfer

onto another object or person. Tertiary and quaternary transfers are also possible. All extra contacts (secondary, tertiary and quaternary transfers) of the item happening before the submission to forensic laboratory have to be considered as a coincidental transfer or natural occurrence.

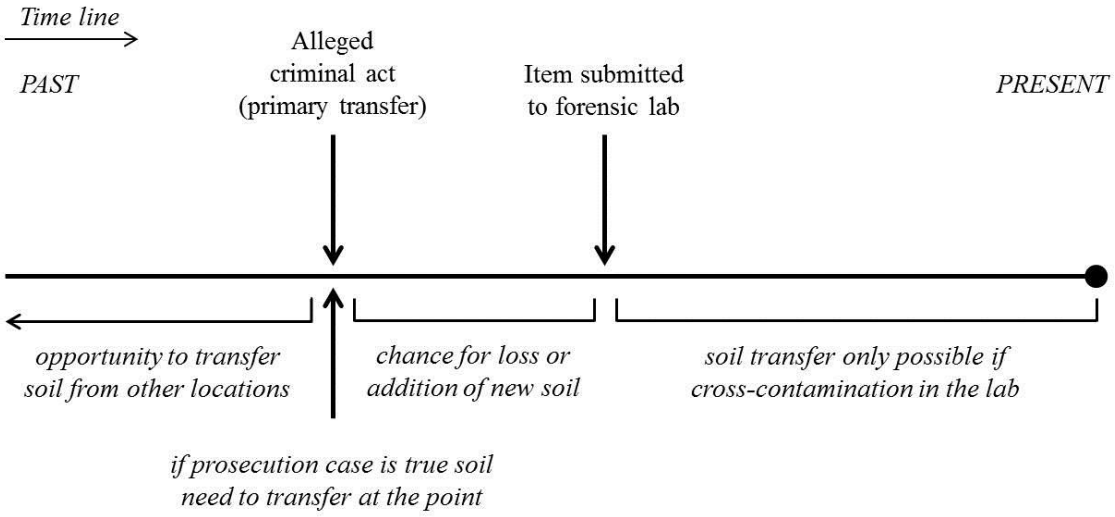



**Figure 1.1. Timeline of the primary soil transfer and possible coincidental transfers.**

The particulate nature of soil material may result in the selective transfer of some particle types onto an item and selective retention after the transfer has taken place (Bull et al. 2006). A further important aspect of forensic soil sample analysis is to be aware of the persistence timeframe of transferred soil trace material. The retention or preservation of soil contact traces that might be present initially are strongly dependent on the nature of the objects that come into contact and the nature of the material exchanged.

Often, in crime investigations there might be a long period of time between the alleged committing of a crime and a suspect being arrested (Pye 2007). Some soil properties may change over this time especially due to the decomposition of organic

components. The analysis of such materials should be undertaken as soon as possible after a crime has been committed, even if questioned samples are not available at that time. Whenever there is a time delay between control sampling and analysis of questioned samples, the potential effect on the results must be factored into the expert opinion.

### 1.2.3 Selection of analysis method and re-analysis

In the forensic context, the word 'trace' commonly means an extremely small amount sometimes even hardly visible with the naked eye (Houck 2004). Here a soil 'trace' sample is considered to be a tiny amount of material in comparison with the control sample available. Therefore, the choice of relevant analytical techniques is determined by the size of the available questioned soil sample. The amount of sample available also dictates the order in which procedures are undertaken. Generally, if a specific chosen technique is destructive to the sample it must be undertaken after non-destructive techniques have been applied.

The primary concern for all methods of soil analysis, particularly in forensic investigations, is the reproducibility of the information gathered. It is recognised that it is better to use fewer methods and to evaluate the reproducibility of the results through a series of different experimental runs rather than to use many techniques with no repeated checks (Morgan & Bull 2007).

Soil samples, taken from a suspect's clothing, footwear or vehicles often represent a mixture of soils from different sources. This anthropogenic nature of soil samples has a considerable impact on the selection of analytical techniques. For example, recent development of automated methods for soil analysis allowed standardisation of the analysis procedure and exclusion of human subjective

decisions. Such automated analysis requires the homogenisation of samples in order to obtain representative composition throughout a sample and its subsequent aliquots. Unfortunately the use of such methods for the comparison between 'anthropogenic'-derived samples and more 'natural' samples (such as those from a crime scene) is unfeasible. Due to the inability of such analysis methods to identify if mixing of soils from different sources has occurred there will always be a potential for false-negative or false-positive interpretations of the study results (Morgan & Bull 2007).

It is also good practice in forensic analysis for at least half of the sample to be left in its original state in order that other scientists may be able to perform sample re-testing if required (Pyrek 2007).

Taking these factors into account it is then possible to prioritise which types of analysis of soils are appropriate and can provide useful information for forensic investigation.

### 1.2.4 Comparison based on exclusion

An important concept underpinning any comparison science is the idea that "two physical objects, in a theoretical sense, can never be identical" (Murray & Tedrow 1991). The inherent nature of soils means that there are no two samples that are precisely the same and even sub-samples of the same soil sample will differ in some manner therefore two forensically relevant soil samples cannot 'match' perfectly (Morgan & Bull 2006). However, it is possible to establish "with a high degree of certainty that the sample is or is not associated with a given scene" (Murray & Tedrow 1991). A fundamental aim in forensic soil examination is to exclude samples from having derived from a common location. In forensic analysis a scientist will make firm opinions of exclusion when two samples are found to be

different in terms of several major aspects at the macroscopic and microscopic levels. Otherwise, it may be concluded that there are no significant ('significant' can have a different meaning in statistics) differences between the samples and therefore the possibility that the questioned sample and control samples are associated in some way cannot be excluded (Pye 2007). If two samples appear to have similar physical, chemical or biological characteristics, they could be derived from one location. At the same time there is the possibility that they could be derived from separate sites having similar characteristics. The difficulty arises in attempting to provide a measure of the significance of the observed similarities and the degree of probability that the questioned sample did originate from the crime scene in order to exclude other locations.

## 1.3 Soil composition and properties

Soil represents an incredibly complex mixture of both living and non-living matter produced as a result of geochemical and biological processes occurring on the Earth's surface. Natural soil properties show large spatial variation that depends on a range of factors, such as geological parent material, climate, topography, biological influences and time (Brady & Weil 2008). In addition, human activity has become one of the most significant factors affecting soil.

The principal constituents of soil are air, water, non-living matter, and various types of living organisms (Figure 1.2) (McCauley et al. 2005). Typically, 45% of soil is non-living matter represented by inorganic mineral particles. Soil often contains assemblages of different particle types that can be divided into the substructures known as aggregates that are classified by size and stability. Further, soils also can be classified by the individual particles or grains the soil samples

consist of. There are numerous types of minerals found in soil. These minerals differ considerably in size and chemical composition. According to mineral particle size, soils can be divided in three classes: sand (2 – 0.05 mm), silt (0.05 – 0.002 mm) and clay (<0.002 mm) (McCauley et al. 2005).

Surface soils usually contain approximate 5% of organic matter which includes: living organisms (soil biota), plant and animal remains and decomposed organic compounds commonly called humus. Soil biota represents a unique symbiosis of living creatures such as bacteria, soil algae, fungi, plant roots and various species of invertebrate and vertebrate animals (Coleman 2001).

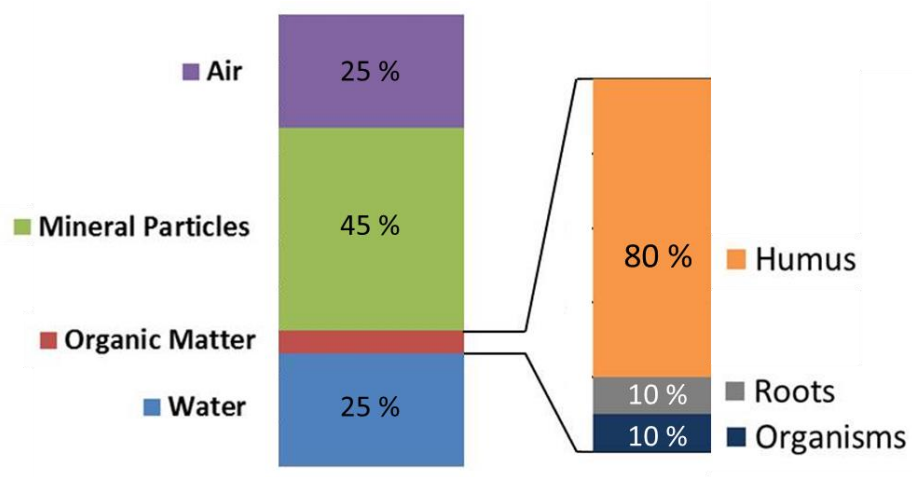



**Figure 1.2. Typical soil composition. The values given above are for an average soil and will vary depending on land-use** (McCauley et al. 2005)**.**

Soil is the largest reservoir of genetic biodiversity of microorganisms on Earth (Delmont et al. 2011). The biological activity of soil is mostly concentrated in the top layer of soil. It accommodates microorganisms from two prokaryotic domains – bacteria and archaea, and some small eukaryotes such as fungi, protozoa and nematodes (Sensabaugh 2009). Prokaryotes dominate the soil environment numerical terms of organisms and species diversity. One gram of soil contains billions of microorganisms of thousands different species (Roesch et al. 2007; Fierer et al.

2007). The most abundant bacteria in soil were found to be Proteobacteria, Acidobacteria, Actinobacteria, Verrucomicrobia, Bacteroidetes and Firmicutes (Janssen 2006). Archaea may constitute up to 10% of all prokaryotes; however this number may vary depending on the soil type (Bates et al. 2011). Fungi can be found in any soils, especially in those that are rich in organic matter where they can make a major contribution to soil biomass (Kennedy & Stubbs 2006). The diversity of fungi in soils can be comparable to that observed for prokaryotes (Fierer et al. 2007). Protozoa are single-celled organisms that feed on bacteria and another organic material in soils. Among the protozoa, the flagellates are the most abundant followed by the amoebae (Esteban et al. 2006). Nematodes are non-segmented worms up to 1 mm in length; they feed on bacteria, archaea, fungi, protozoa, and other nematodes as well as plant and algae material. The abundance of nematodes depends on soil type but do not exceed several hundred per gram of soil (Yeates 2003).

The structure of a soil microbial community is a complex and, more importantly, a dynamic system. Both the chemical and biological characteristics of soil undergo significant changes over different seasons, especially in places where seasons differ sharply. Changes in temperature, humidity, precipitation, natural and man-made disasters can significantly affect microbial community structure.

In summary, it is evident that the highly diverse and unique picture of soil found on the Earth's surface makes soil remarkably valuable and useful in a forensic context. There is no doubt that the real value of soil as an evidential material for forensic investigation could be revealed in full only by a powerful and accurate analytical comparative method capable of distinguishing chemical or biological compositional variation of soil from place to place and over time.

## 1.4 Methods of soil analysis in forensic science

Given soil is a complex and diverse material the use of multidisciplinary descriptive and analytical approaches is necessary to obtain a full and comprehensive picture of its composition. According to the CAFSS guidelines (Figure 1.3), (Fitzpatrick & Raven 2013) there are four main steps in forensic soil examination:

"initial screening of samples, which involves morphological characterisation of bulk soil samples";

"semi-detailed soil characterization, which involves identification, characterisation and quantification of minerals and organic matter in bulk and on individual soil particles following sample selection and size fractionation (< 50 μm or 100 μm)";

"detailed characterization involving additional analytical techniques and/or methods of sample preparation, separation or concentration (e.g. size or magnetic or heavy mineral fractionation) to characterise and quantify minerals and organic matter in bulk and on individual soil particles";

"integration and extrapolation of soil information from one scale to next. It is of note that the analysis of various components of soil could give a different type of information not necessarily matching each other".

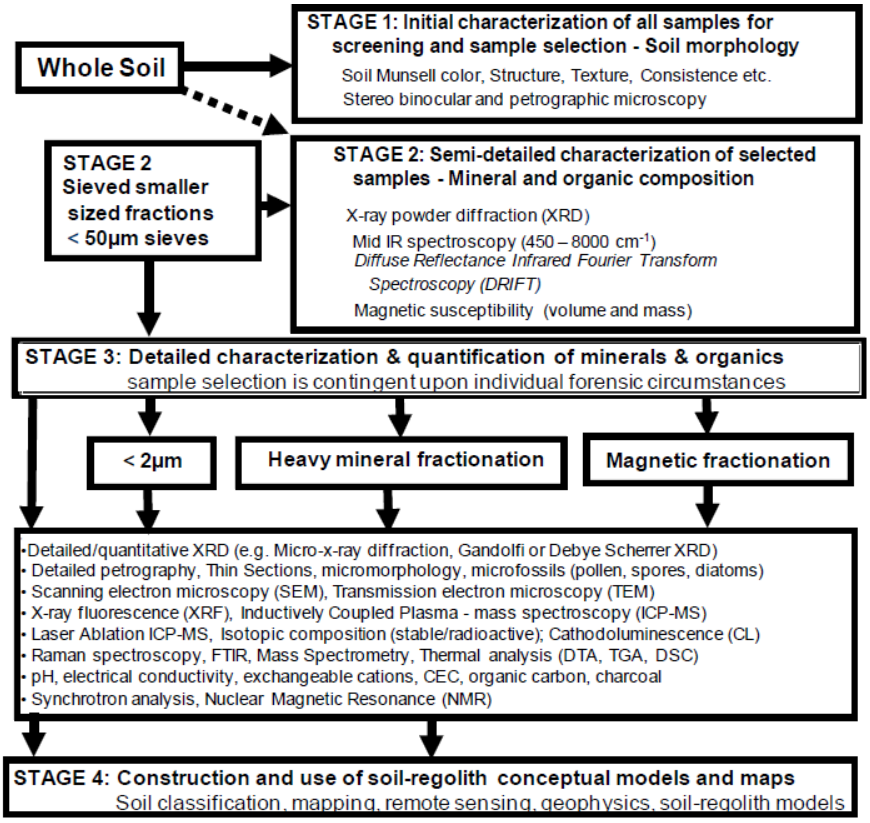



**Figure 1.3. A systematic approach to discriminate soils for forensic soil examinations. (reproduced from Fitzpatrick *et al* (Fitzpatrick 2013) with permission).**

There is a wide range of different physical, chemical, mineralogical and biological parameters to assess in a soil sample. Given that so much of the soil is made of minerals it is not surprising that initial methods of forensic testing of soils have come from geology and mineralogy. Over the past few decades, soil scientists have implemented a variety of modern, sophisticated analytical approaches and tools in criminal investigations. Forensic experts usually apply the most appropriate analytical methods on a case-by-case basis that always requires an element of human perception and judgement.

It has been accepted that forensic soil examination is a challenging task because of the complexity of soil. On the other hand, this diversity and complexity provides an excellent opportunity to characterise, classify and compare soils.

Properties of soil that can be observed or measured directly, and for which significant spatial variations exist, offer the greatest discriminatory value.

### 1.4.1 Soil morphology

Visual examination of soil samples using a simple and non-destructive techniques is one of the first actions performed by a forensic expert after acceptance of the sample. Many of these techniques require minimal equipment such as a low power microscope or the naked eye, and can be performed by any forensic laboratory (Fitzpatrick 2009). Individual particles of both artificial (such as glass fragments, paint chips, weld spatter, rubber pieces) and natural origin (e.g. leaves, broken branches, pollen and spores) undergo extensive investigation and identification and if required are removed for further examination by an appropriate expert (Pye 2007). Morphological soil investigation for assessment of four primary characteristics namely: colour, consistency, texture and structure is then carried out. For instance, soil comparison executed using the Munsell Colour Charts can be a fast, efficient and a cheap starting point, which may give sufficient discriminating power for excluding any similarity between samples (http://munsell.com/about-munsell-color/). This is supported by the study of Morrison *et al.* showing successful colour-based discrimination of soil samples taken from different land-use types within urban areas in the UK (Morrisson et al. 2009). Particle size distribution analysis is performed in a similar way but using unique sets of sieves with standardised mesh sizes. Bonetti and Quarino showed that the particle size distribution method allows for reliable discrimination of soil samples collected from different locations (Bonetti & Quarino 2014). When analysing physical characteristics of soil, the potential for subjective interpretation exists. Also measurement of many of these soil characteristics requires

more than 1 g of soil, which is rarely available in forensic investigations. Moreover, if these soil characteristics are consistent over large distances around the crime site and no distinct and unique features are found, then more detailed methods are necessary.

### 1.4.2 Identification of inorganic and organic components in soil.

*Polarised light microscopy* is a simple and widely used technique performed by forensic soil scientists in order to identify mineralogical composition of a soil sample (Dawson & Hillier 2010). Data obtained during the analysis are typically compared with a standard Michel-Levy chart containing information about interference colours, birefringence and grain thickness that allows for the identification of specific minerals in a sample (Wheeler & Wilson 2008).

*Scanning Electron Microscopy (SEM)* is a very useful approach for forensics because it allows for examination of tiny particles at very high magnification. Manual SEM studies are primarily focused on the detailed three-dimensional imaging of the individual mineral grains rather than bulk soil samples. This type of analysis is highly time-consuming and often depends on an operator's expertise and experience. Manual SEM is a good option for examination of rare, 'exotic' particles providing valuable information on the provenance of a forensic sample (Rawlins et al. 2006). The implementation of the energy dispersive spectroscopy (EDS) along with SEM allows for characterisation of the elemental composition of the mineral particles (Pye & Croft 2004; Pirrie et al. 2014). With the development of automation of this procedure, such as in the QEMSCAN technology, the sample examination became operator independent and enabled detailed characterisation of the mineral composition from just 10 mg of soil sample (Pye & Croft 2004). About 20-30

minerals with the abundance of more than 1% occur in most of soils (Dawson & Hillier 2010). Interestingly, that relative abundances of these dominant minerals often differ significantly in the soils originating from different locations, mainly on a regional scale, which constitutes the basis for SEM-EDS-based mineralogical discrimination of soils (Cengiz et al. 2004).

*X-ray diffraction (XRD)* is one of the most powerful and reliable methods of identifying major minerals and other crystalline structures in soils (Ruffell & Wiltshire 2004). In the United States of America, X-ray diffraction data are accepted as legitimate 'signatures' of the provenance of samples (Dawson & Hillier 2010). The method is based on the specific arrangement of atoms, ions and molecules within a crystalline structure. The sample is analysed by passing X-rays through the sample and measuring the angle of diffracted X-rays that resulted in distinctive X-ray pattern which is unique for each crystalline material (Hubert et al. 2009). XRD requires only few milligrams of soil which is beneficial for forensic investigation however considerable imprecision in the intensity and quality of diffraction patterns can occur due to variation in sample preparation.

*Fourier transform infrared spectroscopy (FTIR)* is a method for obtaining an overall chemical fingerprint of both organic and inorganic compounds found even in a minute amount of sample (less than 1 mg) (Dawson & Hillier 2010). FTIR analysis is a highly precise and sensitive method borrowed from modern analytical science. Cox *et al.* tested this technique on 100 soil samples and showed that FTIR analysis permitted discrimination of soils of similar colour characteristics assessed by the Munsell colour chart (Cox et al. 2000). Woods *et al.* demonstrated that Attenuated Total Reflectance Fourier Transform Infrared spectroscopy (ATR-FTIR) can

efficiently be used as a screening test for discrimination of 'forensic-sized' soil samples prior to submission for more detailed analysis (Woods et al. 2014).

*Elemental composition analysis* belongs to physico-chemical analysis of soil samples. The following range of techniques is frequently used to determine the composition of major trace elements in a soil sample: X-ray fluorescence (XRF), atomic absorption spectroscopy (AAS), inductively coupled plasma (ICP) spectrometry, inductively coupled plasma-optical emission spectrometry (ICP-OES), inductively coupled plasma-mass spectrometry (ICP-MS), neutron activation analysis (NAA) (Pye & Croft 2004). Each technique has its advantages and limitations in terms of sample preparation, size, precision and accuracy. ICP spectrometry has been used in forensic comparison of soils to measure the abundance of a broad range of elements (around 60) in samples as small as 100 mg (Concheri et al. 2011). However, it often requires special sample pre-treatment procedures such as homogenisation, used in order to obtain a representative subsample which is not ideal for forensic investigation with limited sample quantity or multiple secondary soil transfer.

### 1.4.3 Biological materials

Biological material encountered in soil analysis such as plant fragments (Aquila et al. 2014), pollen grains and spores (Morgan, Allen, et al. 2014; Morgan, Flynn, et al. 2014), diatoms (Verma 2013; Pye & Croft 2004; Scott et al. 2014) and microorganisms (Sensabaugh 2009) have been tested for forensic soil characterisation and discrimination.

*Fragments of plant material* such as roots, leaves, stems and seeds can frequently be found in soil by traditional microscopic methods and subsequently

specific plant species may be identified using DNA typing approaches (Aquila et al. 2014; Iyengar & Hadi 2014).

Pollen and spore grains are produced by plants and have distinctive morphological features that allow them to be identified and characterised by microscopic methods. Pollen and spore assemblages present in soils are a valuable source of information relevant to a specific environmental habitat and the surrounding vegetation of a particular site. Analysis of pollen and spores has been presented in court as evidence in a number of forensic cases (Horrocks et al. 1999; Brown et al. 2002; Mildenhall et al. 2006; Walsh & Horrocks 2008). However, their use in the justice system is still under consideration. Objections have been raised that this approach strongly depends on personal opinion, expertise and experience and cannot be generalised across all forensic laboratories (Walsh & Horrocks 2008).

The diversity of *diatoms* found in soils has been also used for forensic investigations using microscopic techniques (Pye & Croft 2004). Diatom frustules are very resistant to decay and often well preserved. Individual diatom species are highly environmentally specific and can be employed as useful natural distinctive features for forensic investigations (Verma 2013).

As already mentioned, *soil* harbours an enormously diverse *microbial community* (Giri et al. 2005). These communities are highly specific to locations, especially those with different soil management and vegetation. It is known that more than 99% of the microorganisms in a soil community cannot be cultured under laboratory conditions, that is why a significant part of the soil microbial community is still considered as ‘unknown’ (Sharma et al. 2014). Given that every organism on Earth has its own unique genome, molecular DNA typing techniques provide a

means for the cultivation-independent analysis of soil community members (Kirk et al. 2004). Analysis of soil DNA material has attracted much attention over the last decades as a useful type of non-human biological evidence (Iyengar & Hadi 2014).

The review of techniques discussed below mainly focuses on the molecular biological DNA typing methods that have been applied for forensic discrimination of soils.

Soil microbial community analysis involves extraction of DNA from the soil sample and then amplification of the DNA material if needed and nucleotide sequence analysis. Considering the limited size of forensic soil samples, most of the DNA typing approaches tested for forensic purposes to date employ an amplification step. The methods for the amplification of total soil DNA extracted can be divided into two groups: (1) gene targeted amplification techniques and (2) random primed amplification techniques or target non-specific techniques (Rincon-Florez et al. 2013).

The favoured markers for the majority of the PCR-based methods are the small subunit (SSU) and large subunit (LSU) of ribosomal RNA operon (16S in prokaryotes, 18S in eukaryotes) residing in the genomes of all living organisms (Tringe & Hugenholtz 2008). Another common target is the internal transcribed spacer (ITS) located between the SSU and LSU of rRNA genes (Anderson & Cairney 2004). These genes contain highly conserved segments of DNA that allow the establishment of phylogenetic relationships amongst very distantly related species. These genes also contain highly variable regions with multiple polymorphisms.

DNA fingerprinting methods, microarrays and molecular cloning have been the principal methods of choice over the last few decades for an assessment of soil biodiversity (Rincon-Florez et al. 2013; Simon & Daniel 2011).

Before the development of high throughput DNA sequencing approaches the most widely used technique in microbial biodiversity analysis was length polymorphism analysis of ribosomal genes (Spiegelman et al. 2005). Many DNA fingerprinting techniques based on the analysis of the length variation of targeted gene regions have been developed for the analysis of microbial community structures. These methods are predominantly based on separation of PCR products by high-resolution gel or capillary electrophoresis.

Amplified ribosomal DNA restriction analysis (ARDRA) has been used as a method for rapid comparison of microbial diversity in a number of environments including soils (Muyzer 1999). After amplification of a particular gene region the resulting PCR products are digested with restriction enzymes and then analysed using a polyacrylamide gel. The disadvantage of ARDRA is that results can be complicated and difficult to interpret, especially in the case of a highly diverse soil microbial community. In the study published by Concheri G *et al*. ARDRA was shown to be successful at the discrimination of microbial DNA fingerprints between two soil samples from suspect and crime scene (Concheri et al. 2011).

The most commonly used DNA-fingerprinting technique in both criminal and environmental applications of soil analysis is terminal restriction fragment length polymorphism (TRFLP) analysis (Liu et al. 1997). TRFLP is a further evolution of the ARDRA technique. In this technique, one of the PCR primers is labelled with a fluorescent dye that allowed the final visualisation of the digested PCR amplification

products with capillary electrophoresis. The use of TRFLP as a forensic soil comparison method was proposed by Horswell *et al.* (Horswell et al. 2002). The authors demonstrated that they were able to generate microbial DNA fingerprints from small soil samples recovered from shoes and clothing. Subsequent profile comparison indicated that soil samples taken from the same location had a higher degree of similarity than those from different sites. This method was especially promising for forensic applications because capillary electrophoresis instrumentation used for DNA fragments separation is ubiquitous in forensic DNA laboratories (Heath & Saunders 2006) Lenz & Foran, 2010 (Macdonald et al. 2008) (Macdonald et al. 2011). Profile interpretation after TRFLP may be problematic as fragments with the same length can be different in sequence yet migrate to the same position generating false positive results.

Length-Heterogeneity PCR (LH-PCR) produces characteristic soil metagenomic DNA profiles based on the sequence length hyper-variability existing within the 16S rRNA genes or inter-genus spacer region (Spiegelman et al. 2005). Similar to TRFLP, the LH-PCR method uses fluorescently labelled primers set for PCR amplification. Resulting PCR-products then can be directly analysed by capillary electrophoresis. Limitations of LH-PCR are the same as for TRFLP, which include inability to resolve complex DNA mixtures and the possibility of misleading interpretation as phylogenetically different species may have DNA fragments of the same length. Nevertheless, recently this method was shown to be robust and reproducible for the monitoring of the soil microbial community changes associated with cadaver decomposition (Moreno et al. 2011).

A different principle of PCR-products separation was introduced in Denaturing Gradient Gel Electrophoresis (DGGE) and Temperature Gradient Gel

Electrophoresis (TGGE). These techniques use polyacrylamide gel separation of amplified ribosomal DNA fragments of the same length but different in nucleotide composition (Muyzer & Smalla 1998; Spiegelman et al. 2005). These approaches have also been applied to forensic soil comparison (Lerner et al. 2006; Pasternak et al. 2012) . It is not always feasible to separate similar fragments with different sequences because of the similar thermodynamic stability of the fragments. As a result the interpretation of DGGE and TGGE analyses can often be misleading since a single electrophoretic band may be derived from multiple species. Though, these methods allow for each band to be excised from the gel and subsequently cloned and sequenced revealing phylogenetic characterisation of microbial community members. Until recently preliminary microbial diversity surveys were carried out using clone libraries despite being a labour-intensive and time-consuming process (DeSantis et al. 2007; Janssen 2006). Assessment of actual bacterial diversity in soil using these techniques is problematic, because it has been shown that to document 50% of the species richness in soil 40,000 clones of the 16S rRNA genes are required (Rastogi & Sani 2011).

Another approach for the assessment of total soil DNA composition is a randomly amplified polymorphic DNA (RAPD) analysis (Franklin et al. 1999). In standard PCR primers are used to amplify known/target regions of the DNA. In RAPD a single randomly chosen primer or set of primers is used to amplify DNA material of unknown sequence. Arbitrarily primed PCR (AP-PCR) (Welsh & McClelland 1990) and DNA amplification fingerprinting (DAF) (Caetano-anollds et al. 1991) techniques were independently developed methodologies. These methods differ in primer length, amplification stringency and procedure used to resolve DNA patterns (Caetano-Anolles 1993). The major drawback of the methods is

amplification conditions which may vary between two different laboratories and therefore significantly influence resulting fingerprints (Tyler et al. 1997). Waters *et al.* proposed to assess soil DNA content by hybridisation of the obtained AP-PCR amplification products with custom microarrays. The approach was shown to be successful at generation of reproducible and discriminatory soil DNA profiles (Waters et al. 2012).

## 1.5 Metagenomics as a new tool for forensic soil discrimination.

The assemblage of all genetic material from all microorganisms residing in soil is called metagenome (Handelsman et al. 1998). The main concept of metagenomics is that all organisms could be identified through DNA analysis in given environmental samples without culturing (Handelsman 2004). Metagenomic approaches allow investigation, classification and manipulation of the entire genetic material isolated from a sample (Tringe & Rubin 2005; Fuhrman 2012).

The basic steps of metagenomic analysis are: ***a)*** the isolation of genetic material directly from the environmental samples; ***b)*** manipulation of the genetic material (which may, or may not, include amplification of DNA material and library preparation); ***c)*** the analysis of genetic material in the metagenomic library; and ***d)*** bioinformatics data analysis.

***a)*** Many different approaches for the extraction of total DNA material from environmental samples have been developed (Roose-Amsaleg et al. 2001). The extraction process can result in specimens that do not contain an even representation of all organisms present in a sample (Feinstein et al. 2009). Available commercial

kits, tested for different soil types, represent a mean for standardisation of the DNA extraction procedure and allow for inter-laboratory comparisons to be performed. Most of the soil DNA extraction kits include steps of homogenisation of the soil particles, chemical lysis of microbial cells and absorption of the released nucleic acid on silica sorbents (silica membranes in spin column approaches or silica coated magnetic beads) in the presence of chaotropic agents (such as guanidine hydrochloride, guanidine thiocyanate or potassium iodide) in high concentrations (Boom et al. 1990). Following washing of the sorbent with 70% ethanol helps to remove excess of salts and prepares nucleic acid to be eluted with low ionic strength buffers, such as 10 mM Tris-EDTA solution or water (Azad et al. 1991; Boom et al. 1992).

***b)*** There are two ways to manipulate extracted DNA material: (a) total genomic DNA from different organisms is cut up into small length fragments using enzymes called restriction endonucleases; or (b) PCR amplification of taxonomically informative markers can be performed. As already mentioned, SSU and LSU rRNA loci (for example, 16S, 18S, ITS) are the most widely used markers in microbial community structure (Hajibabaei 2012). The obtained random DNA fragments or PCR products are then cloned into a chosen bacterial vector system, typically based on the *E. coli* replication system, and screened using Sanger sequencing (Shendure & Ji 2008). Sanger sequencing technique is based on PCR amplification with dideoxy nucleotide triphosphate terminators followed by CE separation of the obtained PCR products for each clone (Sanger 1977). This technique was a dominant sequencing platform until the release of the first NGS technology in 2005 and has been applied in most of the DNA-based studies of environmental microbial communities (Tringe & Hugenholtz 2008). The major limitations of this method are the high cost and time

required as well as the phenomenon whereby only a few hundred clones capture mostly dominant members of the microbial community, leaving the contribution of low-abundant species undervalued (Hur & Chun 2004).

Development of high-throughput DNA sequencing (HTS) technology allowed for the independent sequencing of random DNA fragments (shotgun sequencing) or PCR products (targeted 16S rRNA sequencing, for example) directly from environmental samples that in turn revolutionised metagenomic studies. However, HTS platforms require special DNA pre-processing procedure. In the shotgun approach, high molecular weight DNA is firstly fragmented by mechanical shearing or enzymatic digestion into an appropriate platform specific size range followed by enzymatic adapters ligation (Buermans 2014). In the case of PCR-based targeted sequencing (16S sequencing), PCR primers can be designed in such a way to contain adapter sequences. In both cases DNA samples can also be tagged with specific barcodes (short oligonucleotide sequences) that allows for simultaneous analysis of multiple samples. The emergence of HTS approaches has yielded a greater insight into microbial diversity of complex environments, such as soil, and permitted monitoring of the spatial and temporal variation of microbial communities of different soil types (Roesch et al. 2007; Lauber et al. 2013; Wang et al. 2013).

***c)*** The first platform for high-throughput DNA sequencing, the 454 GS 20 pyrosequencing platform, was developed by Roche (Roche Diagnostics Corp., Branford, CT, USA) in 2005. The pyrosequencing method revolves around detection of the pyrophosphate released during primer extension reaction. The released pyrophosphate triggers a luciferin to oxyluciferin conversion reaction cascade, which is accomplished with a chemiluminescence light emission. The reaction occurs on micro-beads deposited in a picotiter plate (Margulies 2005). The output of the

sequencing process is a number of short DNA sequences also called reads. Initially the 454 platform generated up to 10 Mbp per run with an average read length of 100 bp. The latest versions of the pyrosequencing platform from Roche were the GS FLX+ and GS Junior+ that produce 700 and 70 Mb of information per run with a read length of 700 bp (http://www.454.com/). It is of note that the pyrosequencing platform was announced to be abandoned from mid-2016.

In early 2007 Illumina released the Solexa platform that evolved into the MiSeq, HiSeq 2000/2500, HiSeq X Ten and the NextSeq 500 platforms. The Illumina platform uses reversible terminator chemistry where each deoxynucleotide triphosphate is blocked at 3' hydroxyl with a photocleavable fluorescent dye. Initially denatured DNA fragments with adapter sequences hybridise to surface functionalised primers (complementary to the adapters) in a flow cell, which further allows for surface polony bridge amplification. During each cycle one nucleotide is incorporated. Then, after surface imaging the fluorescent 3' blocker is cleaved off and new incorporation step is repeated (Bentley 2008). HiSeq 2000/2500 provides up to 1000 Gb of sequences per run with an average length of 150 bp, where the benchtop version MiSeq is able to produce up to 15 Gb of sequences of 300 bp length in average (http://www.illumina.com/applications/sequencing.html).

Life Technologies (which was acquired by Thermo Fisher Scientific in February 2014) has two NGS platforms: SOLiD System and Ion Torrent (https://www.lifetechnologies.com/au/en/home/life-science/sequencing.html).
SOLiD sequencing technology employs multi-round dibase incorporation system by ligation (Peckham 2007). SOLiD 5500XL Genetic Analyser generates up to 10-15 Gb reads per run with a length of 75 bp. The Ion Torent technology is not based on the imaging of fluorescent/chemiluminescent signals but on release and sensing of

the $H^+$ (hydrogen ion) during solid-phase PCR amplification. This occurs in the specialised Ion Torrent chip consisting of millions of micro-wells with imbedded pH sensors. Each well is filled with a sequence template anchored to a microbead. Four nucleotides are washed through the chip in a consecutive order that results in the pH change of the microwell environment if the nucleotide incorporation occurs (Buermans 2014). Currently the Ion PGM system can, using the latest Ion 318 chip, produce up to 6 million reads of 400 bp in length, or 2 Gb per four-hour run. The newer Ion Proton platform with the Ion PI chip provides up to 80 million reads with an average length of 200 bp in the same time frame, or 10 to 14 Gb per run.

Other existing sequencing platforms are PacBio RS SMRT system (Pacific Bioscience, Menlo Park, CA, USA) and GridION/MinION (Oxford Nanopore Technologies , UK). They allow single-molecule sequencing with much longer read length of up to $10^4$ bp (Pennisi 2014).

***d)*** A large number of different systems and resources for bioinformatics analysis of HTS data were developed in the form of on-line web portals, web services and stand-alone programs. There are two main approaches for characterising sequencing data (Mande et al. 2012a). In the first approach, also referred as 'taxonomy dependent analysis', individual reads are taxonomically classified by comparing them to sequences of known phylogenetic origin available in such public repositories as NCBI (http://www.ncbi.nlm.nih.gov/), EMBL (http://www.embl.org/) and DDBJ (http://www.ddbj.nig.ac.jp/). Similarly, functional annotation via mapping reads to protein libraries (non-redundant databases) can also be performed. For the taxonomic or functional assignment of reads alignment-based algorithms such as Basic Local Alignment Search Tool (BLAST) (Altschul et al. 1990) and BLAST-like alignment tool (BLAT) (Kent 2002) adopted by the major annotation pipelines as

Metagenomic Rapid Annotations using Subsystems Technology (MG-RAST) (Meyer et al. 2008) and Integrated Microbial Genomes (IMG/M) (Markowitz et al. 2012) are often used which result in generation of metagenomic abundance profiles of phylogenetic and metabolic features. However, in case of the highly diverse microbial communities, for example like soil, characterised by a large number of previously unknown organisms, an assignment of individual reads by taxonomy-dependent method is often unfeasible. In other words, sequence datasets from such complex metagenomes contain reads that have no 'genomic reference' and consequently cannot be mapped to the known 'taxonomic reference tree' (Mande et al. 2012b) . Usually such reads are allocated in the group as 'unassigned' or 'unknown'.

In contrast, 'taxonomy-independent' approach simply groups or compares reads in given datasets based on their nucleotide sequences similarity (Mande et al. 2012b).

Metagenomic profiles contain cosiderably larger amount of taxonomic or metabolic features (variables) than the number of samples analysed, therefore a range of multivariate statistical analyses including the hierrahical agglomerative clustering (CLUSTER) (Jain et al. 1999), non-metric multidimentional scaling (NMDS) (Clarke & Green 1988), analysis of similarities (ANOSIM) (Clarke & Green 1988) are often employed for the assessment of differences between samples.

While the metagenomic sequencing of soil community has been already widely applied to ecological and environmental studies, the potential of this approach in the forensic field has not yet been fully evaluated.

## 1.6 Forensic validation of new approaches

The High Court in Australia examines Supreme Court decisions made in other jurisdictions but is not bound by them. The admission of a new forensic technique into the Australian courts is influenced by judgements made in U.S. courts but relies more on admissibility and precedent. The US supreme court gave direction to the U.S. courts for the acceptance of new expert scientific evidence into the courtroom following the case of Daubert v Merrell Dow Pharmaceuticals (509 US 579 (1993)) and its subsequent appeal.

The directive firmly places the Judge as the gatekeeper for determination of whether a new technology is admissible in a court of law and reliability and relevance are of prime consideration.

The Daubert judgement gave a set of criteria that are to be considered by the Judge in determining the admissibility of scientific evidence at trial:

- Whether the scientific theory or technique can be and has been tested (validated)
- Whether the underlying scientific theory or technique has been subjected to peer review and publication
- Whether the scientific theory or technique has a known or potential rate of error, and if standards exist to control the technique's operation and error
- Whether the scientific theory or technique has attained general acceptance within the relevant scientific community (not discipline as per Frye (Frye v, United States 293 F 1013 (1923)).

The Daubert criteria are applied to varying degrees across the US and a strict application of the criteria turns very much on the individual Judge hearing a trial. Generally there is a stricter testing of new technologies in American courts than in Australian courts. The newer technologies, such as DNA in R v Karger (2001), have undergone relatively stringent testing prior to acceptance in Australian courts but the 'Police sciences' such as fingerprints and tool marks have a long history with little challenge. These sciences are coming under increased scrutiny in Australian courts and their conformance with Daubert is becoming questioned (Edmond & Mercer 2004). So much so that a range of Special Working Groups and Special Advisory Groups have been formed within each of the disciplines in order to establish good science practices for existing forensic analytical techniques.

The introduction of any new technology into the Australian court system will be best-suited to fit within the Daubert criteria and this would also make it more amenable to be translated to international courts. This includes: (a) effective planning and conducting of scientific proof-of-concept studies that are designed to demonstrate the potential of a new forensic technique. This stage is primarily focused on revealing the fundamental aspects of the method being performed at standard conditions with few sources of variation and error introduced. Proof-of-concept studies play a crucial role in developing new analytical techniques by helping to understand the strengths and weaknesses of the proposed approach. Based on the obtained results a decision of whether the technique has the potential to fulfil requirements of the Daubert criteria for admissibility in court is made. Moreover, proof-of-concept studies have to be supported by publications in peer-reviewed scientific journals.

Following this a standard operating protocol (SOP) is created. The use of any technology within a research setting is often open to the individual researcher to amend or change to optimise the technique for a given sample or circumstance. Within a forensic application techniques are more ridged and often an analytical technique will be re-validated and assessed after even minor changes. The following requirements apply to a laboratory involved in routine forensic investigation:

- It should have a documented procedure for each analytical technique used;

- It should have a documented approach for testing general unknowns;

- Procedures should include a list of equipment and reagents, step-by-step instructions, quality controls, test calculations, limitations, interpretation criteria, and literature references;

- The laboratory should have a policy whereby a deviation from the analytical procedure is documented and approved (Pyrek 2007).

Final validation studies are designed to demonstrate performance of the developed methods, identify error rates and statistical uncertainties. Commonly these studies are performed on a large selection of forensically representative samples. Additionally, these studies must be run with the SOP to be implemented into active casework and a number of samples will be encrypted so that the laboratory does not have the answer pre-determined. Validation studies amongst other things will assess the specificity, sensitivity, reproducibility, bias, precision, false-positives, false-negatives and determine appropriate controls.

# Thesis Objectives

The ultimate aim of the study presented in this thesis was to evaluate the ability of modern methods of metagenomic analysis of microbial communities for the discrimination of forensically relevant soil samples.

Specifically the aims were:

- the testing of ‘gold standard’ metagenomic approaches such as targeted metagenomics (sequencing of microbial marker genes) and random whole metagenomic sequencing (shotgun sequencing and WGA-based sequencing) widely used in ecological research
- the evaluation of new single arbitrary primed PCR based sequencing approach
- the assessment of important aspects of soil metagenomic analysis such as the quality of reagents (DNA extraction kits and commercial DNA polymerases) and estimation of the impact of different storage conditions on the soil DNA material
- the consideration of various bioinformatic pipelines for treatment and analysis of the sequence datasets representing an inherent part of current metagenomics in order to provide an appropriate interpretation of the sequencing data obtained.

# Thesis structure.

The research is presented in the following chapters:

**Chapter 2** – focuses on the selection of an appropriate soil DNA extraction kit and DNA polymerase for PCR amplification, and the evaluation of the effect of various soil storage conditions on the subsequent metagenomic DNA analyses.

**Chapter 3** – investigates the performance of 16S rRNA targeted metagenomics at discriminating soils taken from three distinct urban sites 5 km apart, two of which were visually similar and from similar environments.

**Chapter 4** – introduces and assesses a new approach based on the arbitrarily primed PCR amplification for the random whole metagenomics study of soil microbial communities and the discrimination of the same soils analysed in Chapter 3.

**Chapter 5** – compares the performance of the shotgun and WGA-based sequencing approaches with AP–based sequencing and their ability to differentiate visually similar soils taken from two urban sites of similar land use and vegetation type 5 km apart (samples used are the same analysed in Chapters 3 & 4).

**Chapter 6** – describes the application of reference independent bioinformatic algorithms for the comparison of the metagenomic datasets from Chapter 5.

**Chapter 7** – summarises the results obtained in previous chapters and provides recommendations to be used for introducing modern metagenomic approaches as a new tool for forensic soil discrimination.

# Chapter 2. Evaluation of soil DNA extraction, amplification and storage impacts on the soil microbial community DNA typing

## 2.1 Introduction

Soil metagenomics involves isolation of total soil DNA material followed by its subsequent analysis using high-throughput DNA sequencing (HTS) techniques. The reliability of any results obtained depends greatly on the sensitivity, efficacy and consistency of the processes involved in the entire analysis. All DNA-based molecular methods of soil microbial community analysis rely on the quantity and quality of DNA specimens. Efficient DNA extraction from soil is an essential step in achieving accurate, reproducible and reliable results of soil microbial profiling using modern metagenomic approaches. Hence, the influence of any relevant factors that could have a substantial effect on the efficiency and consistency of DNA recovery from the soil sample must be considered.

The extraction of high-purity DNA from soil is often a challenging task because of humic acids that are easily co-extracted with soil DNA and may interfere with downstream procedures and applications. During the last three decades, many efforts have been devoted to developing and optimizing soil DNA extraction in order to obtain high-quality metagenomic DNA suitable for characterisation of the whole microbial community (Terrat et al. 2012; Sagar et al. 2014). These efforts led to the development of numerous in-house DNA extraction protocols (Miller et al. 1999; Williamson et al. 2011; Fatima et al. 2011; Zhao & Xu 2012; Taberlet et al. 2012) and more recently commercially available kits (Whitehouse & Hottel 2007; Vishnivetskaya et al. 2014; Knauth et al. 2013). The availability of commercial kits has made soil extraction quick and a straightforward process. The use of commercial kits also represents a means to standardise soil DNA extraction procedures and aids in the comparison of data between laboratories. A review of soil DNA extraction kits and general considerations for their

use in forensic practice has recently been published (Young, Rawlence, Weyrich, & Cooper, 2014).

A number of comparative studies were performed to demonstrate advantages and limitations of different soil DNA extraction kits by assessment of the yield and quality of the extracted DNA (Dineen et al. 2010; Knauth et al. 2013). Also, the obtained soil DNA extracts were further evaluated using various applications including qPCR (Olson & Morrow 2012), microarray applications (Delmont et al. 2011; Ning et al. 2009), DNA fingerprinting analyses (Leckie 2005; Zhao & Xu 2012; Knauth et al. 2013) and metagenomic DNA sequencing (Feinstein et al. 2009a; Taberlet et al. 2012).

The most widely used method for metagenomic analysis of soil microbial communities involves a PCR amplification stage. It is known that a well-optimised PCR can detect just a few DNA molecules (Khodakov et al. 2008), which, in turn, imposes specific requirements for sample contamination monitoring, especially when samples with low quantity of DNA are amplified. Avoiding sample contamination from exogenous DNA must be one of the highest priorities in any PCR laboratory, and particularly in forensic science, where analysis of trace evidential material is commonly performed. This can be achieved by using appropriate equipment (e.g. gloves, facemask, dedicated extraction and PCR hoods, and sterile lab-ware) and strictly following good laboratory practice guidelines. Moreover, assessment and prevention of exogenous contamination is essential for targeted PCR-based metagenomics where gene-specific broad-range universal primers are typically used for the analysis of microbial genetic markers. This issue is associated with traces of microbial nucleic acids found in brand-new reagents, for instance commercially available Taq DNA polymerases (Mennerat & Sheldon 2014). It has been reported previously that commercial DNA polymerase preparations inevitably contain bacterial DNA traces, retained after the production of

the polymerase (Spangler et al. 2009; Mühl et al. 2010). Also, there were recent reports that some DNA extraction kits, and even plastic-ware, might also be initially contaminated with microbial DNA (Young et al. 2014). The preliminary assessment of microbial DNA contamination in available DNA extraction and amplification reagents is therefore necessary for obtaining reproducible and reliable results of metagenomic research.

The proper handling and treatment procedures of soil samples after their collection are also critically important for soil metagenomics research. It is clear that after collection of a soil sample, all natural processes related to the soil microbiome activity continue. It is widely assumed that storage of samples at room temperature, even for a short time period, can make the samples unfit for downstream DNA-based analysis due to changes in the microbial community structure (Lauber et al. 2010). According to the International Organization for Standardization, it is highly recommended therefore to perform the extraction of microbial DNA from the soil samples immediately after collection (Petric et al. 2011; Terrat et al. 2014). However, such immediate DNA extraction is not always feasible and in this case the best practice recommended is to keep the samples frozen (Wallenius et al. 2010). For example, soil forensic evidence samples may have been stored for prolonged periods before analysis procedures commence. Usually in forensic applications, these samples are preserved by air-drying; desiccation has been found to be sufficient for the stabilisation of physical-chemical composition of soil (Dawson & Hillier 2010; Fitzpatrick & Raven 2013). It remains though unclear what kind of storage conditions (chemical preservatives, temperature and duration of storage) would be acceptable for preserving the initial soil microbial composition such that the subsequent DNA-based typing is unbiased. Only a few and controversial results have been reported on the assessment of how storage

conditions influence soil DNA profiles. It has been shown that genetic material in soil is resistant to storage effects and air-drying can be used as cheap and simple method for conservation of the samples (Klammer et al. 2005). Tzeneva *et al*. observed significant effect of air-drying and prolonged storage at room temperature on the soil bacterial composition using the DGGE fingerprinting method but not on the eukaryotic soil community (Tzeneva et al. 2009). Similar results of no effect of air-drying on microeukaryotic soil diversity were obtained later (Zhao & Xu 2012). Another study, based on length heterogeneity PCR technique (LH-PCR), demonstrated that -20 °C freezing retains the structure of the soil bacterial community better than the air-drying method (Wallenius et al. 2010). Rissanen *et al*. successfully used LH-PCR method to show that treatment of a soil sample with a phenol-chloroform mixture allowed for efficient preservation of the sample, which might be useful when freezing of the sample is not possible, e.g. for in-field studies (Rissanen et al. 2010). In contrast, recent studies performed by Lauber *et al*. (Lauber et al. 2010) and Rubin *et al.* (Rubin et al., 2013) showed that characterisation of soil bacterial community structures by means of high-throughput DNA sequencing was not affected by different storage conditions, such as temperature and duration of storage. These studies illustrate that many questions regarding how soil samples should be collected and stored are unresolved and of current interest in forensic science.

The aim of this chapter is to find optimal conditions and regents for performing efficient soil DNA extraction and amplification. In order to achieve the aim the following tasks were designed and executed:

- Collection of soil samples from three different urban locations in Adelaide, followed by DNA extraction

- Testing of different commercial soil DNA extraction kits for their ability to isolate high quality soil DNA preparation suitable for consecutive metagenomic applications. The isolated DNA to be tested for the following parameters: DNA yield, DNA purity and richness of LH-PCR profiling of the collected soil samples
- Testing of commercial DNA polymerases for the presence of trace amounts of bacterial DNA retained after the polymerase production
- Assessment of the soil storage conditions impact effect on the DNA composition of the soil samples by LH-PCR method.

## 2.2 Materials and methods

### 2.2.1 Soil sample collection

Soil samples were collected from three different parkland sites in Adelaide separated by approximately 3 km: location A (S35 01 43.42 E138 34 16.26), location B (S35 00 58.09 E138 32 12.03) and location C (S35.021317 E138.515922). The selected locations represented areas highly disturbed by human activity, such as the footpath in an open space area in the park, playground and near to the sidewalk along the beach. One soil sample (approximately 3 g) for each location was taken from the 2-3 cm of the top layer and placed in individual sterile plastic tubes and stored at -20 °C until it was used for soil DNA extraction kits evaluation. The second set of three fresh soil samples was taken from the same sites at different time point for the soil storage evaluation experiments. The soil samples collected from location A and B represented a dark loam rich in organic matter and were visually very similar (Figure 2.1). The sample from location C represented a sandy soil and could be easily distinguished from the first two ones (Figure 2.1).


**Figure 2.1. Photographs and relative disposition of locations A, B and C and the corresponding soil samples taken.**

### 2.2.2 DNA extraction kits evaluation

Each soil sample was mixed thoroughly and then divided into aliquots of 0.05 g. Five commercially available soil DNA extraction kits were tested:

1. PowerSoil DNA Isolation kit (PS) (MO BIO Laboratories Inc., Carlsbad, CA, USA),

2. UltraClean Soil DNA Isolation kit (UC) (MO BIO Laboratories Inc., Carlsbad, CA, USA),

3. EZNA Soil DNA kit (EZ) (Omega Bio-Tek Inc., Norcross, GA, USA),

4. ZR Soil Microbe DNA kit (ZR) (Zymo Research Corporation, Irvine, CA, USA),

5. FavorPrep Soil DNA Isolation kit (FP) (Favorgen Biotech Corporation, Taiwan)

Extraction of metagenomic DNA from five replicate aliquots of each soil type was performed following to the manufacturers' instructions for each kit. Extraction of blank controls for each kit was performed along with the soil samples processing.

### 2.2.3 Soil processing for storage experiments

Fresh bulk soil sample from each collection site was divided into aliquots of 0.05 g (n = 14). Then a half of aliquots (n = 7) for each soil type was treated with an equal volume of 2-propanol (IPA) and dried under reduced pressure in a desiccator. One aliquot from each subset (i.e. with and without 2-propanol treatment) was used in the DNA extraction procedure using ZR Soil Microbe DNA kit (ZR) within 24 h after the soil samples collection. The remaining aliquots from each subset (n = 6) of each soil type were stored at room temperature ('bench-top storage'), +4 °C ('fridge storage') and -20 °C ('freezer storage'). Total DNA was extracted after two and four weeks of storage after the initial DNA extraction.

In all DNA extraction experiments, the total DNA obtained was visualised by electrophoresis after separation in a 1% agarose gel stained with ethidium bromide. The DNA purity indices of the DNA extracts, such as A260/280, A260/230, A320 and A340, were quantified by UV-VIS spectroscopy using NanoDrop-1000 (ND-1000, NanoDrop Technology, Wilmington, DE, USA). DNA concentrations of the soil extracts were determined using a Qubit dsDNA HS Assay Kit (Invitrogen, USA) on a

Qubit 2.0 fluorometer (Life technologies, USA). Data are presented as an average value ± SD (Standard Deviation) across five replicate extracts obtained by each DNA extraction kit for each soil type.

### 2.2.4 Selection of DNA polymerase

Eight commonly used Taq DNA polymerases (listed in Table 2.1) were selected to be tested for the presence of bacterial DNA traces. PCR amplification was performed with two sets of primers that are specific to bacterial 16S rRNA gene (Table 2.2).

**Table 2.1. DNA polymerases used in the study.**

| *DNA polymerase* | *Manufacturer* | *DNA polymerase* | *Manufacturer* |
|---|---|---|---|
| One Taq | NEB | MyTaq | Bioline |
| Q5 Hi-Fi | NEB | HotStar Taq | Qiagen |
| Taq | NEB | AmpliTaq Gold 360 | Life Technologies |
| GoTaq Flexi | Promega | Mango Taq | Bioline |

**Table 2.2. Sequences of primers for amplification of bacterial 16S rRNA gene used in the study.**

| *16S rRNA region* | *Primer naming* | *Primer sequence* | *Ref.* |
|---|---|---|---|
| V1-V2 (LH) | 27-F<br>355-R | 5'-6FAM*-AGA GTT TGA TCM TGG CTC AG-3'<br>5'-GCT GCC TCC CGT AGG AGT- 3' | (Moreno et al. 2006; Di Bonito et al. 2013) |
| V3-V5 (cc-cd) | cc-F<br>cd-R | 5'-CCA GAC TCC TAC GGG AGG CAG C-3'<br>5'-CTT GTG CGG GCC CCC GTC AAT TC -3' | (Rudi & Larsen, 1997) |

* 6FAM = 6-Carboxyfluorescein

Ten replicates of no-template controls along with two replicates of positive controls of *E. coli* DNA (Sigma-Aldrich, USA), and using 1 ng per reaction, were performed for each polymerase. PCR amplification components and the reaction conditions are shown in Table 2.3. PCR amplification products were visualised by electrophoresis after separation in a 2% agarose gel stained with ethidium bromide.

**Table 2.3. Composition and conditions of PCR amplification reactions used in DNA polymerase selection experiment.**

| | PCR Components | | | | | | PCR Thermocycling | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Buffer µL | $Mg^{2+}$ mM | dNTPs mM each | DNA Pol units | Primer, µM each | DNA µL | Initial denaturation | Denaturation | Primers annealing | Primers extension | Final extension |
| **AmpliTaq Gold 360 DNA Pol (ABI)** | 2.5 (10×Buff) | 2.5 | 0.2 | 0.625 | 0.2 | 1 | 95°C, 15 min | 95°C 30 sec | 68°C (ccF-cdR) 55°C (27F-355R) 30 sec | 72°C 45 sec | 72°C 3 min |
| **HotStarTaq DNA Pol (Qiagen)** | 2.5 (10×Buff) | 2.5 | 0.2 | 0.625 | | | 95°C, 2 min | | | | |
| **MyTaq DNA pol (Bioline)** | 5 (5×Buff) | 3* | 0.25* | 0.5 | | | 95°C, 2 min | | | | |
| **Mango Taq DNA pol (Bioline)** | 5 (5×Buff) | 2.5 | 0.2 | 0.5 | | | 95°C, 2 min | | | | |
| **GoTaq (Promega)** | 5 (5×Buff) | 2.5 | 0.2 | 0.625 | | | 95°C, 15 min | 95°C 15 sec | | | |
| **One*Taq* DNA pol (NEB)** | 5 (5×Buff) | 1.8 | 0.2 | 0.625 | | | 95°C, 2 min | | | | |
| **Taq DNA Pol (NEB)** | 5 (5×Buff) | 2 | 0.2 | 0.625 | | | 95°C, 2 min | | | | |
| **Q5 Hi-Fi DNA pol (NEB)** | 5 (5×Buff) | 2* | 0.2 | 0.5 | | | 95°C, 2 min | | | | |

***** = included in the corresponding buffer. In each case water was added up to make the final PCR volume of 25µL.

## 2.2.5 LH-PCR profiling

LH-PCR amplification was performed using 1×HotStar buffer (Qiagen, VIC, AU), 2.5 mM $MgCl_2$, 0.5U Hotstar Taq DNA Polymerase (Qiagen), 250 µM dNTPs (Promega, USA), 0.2 µM 27-F and 0.2 µM 355-R primers (Table 2.2), DNA template and water to a final volume of 25 µL. As a template, 10 ng of DNA was used as has been reported earlier (Ritchie, 2000). Amplification was performed using the following parameters: initial denaturation at 95 °C for 15 min, 30 cycles of denaturation at 94 °C (30 s), annealing at 55 °C (30 s) and extension at 72 °C (45 s), and a final elongation at 72 °C for 3 min. LH-PCR amplification was performed in triplicates for each soil DNA extract.

## 2.2.6 Capillary Electrophoresis

Analysis of the obtained LH-PCR products was performed using capillary electrophoresis (CE) by adding 0.5 µL of the LH-PCR reaction to 9.5 µL of a 96:1 (vol:vol) mixture of Hi-Di formamide and GeneScan 500 LIZ size standard (Applied

Biosystems). The mixture was then heated for 3 min at 95 °C and snap cooled for 5 min. CE analysis was performed on an ABI 3130*xl* Genetic Analyzer (Applied Biosystems) equipped with a POP-4 polymer (Applied Biosystems) filled capillary array, G5 matrix. Output data were analysed using GeneMapper, v. 3.2 (Applied Biosystems). The minimum noise threshold was set to 50 RFUs (Moreno et al. 2011; Di Bonito et al. 2013).

### 2.2.7 Statistical analysis of LH-PCR profiles

The profiles of relative peaks intensity for each sample were imported into MS Excel Software to be filtered and normalized. The first filtering criterion was that a peak observed in a particular sample to be scored as a true peak had to be present in all three replicate DNA analyses of this sample. The relative intensity ($I_i$) for each PCR-product peak was calculated by dividing the peak intensity ($\bar{h}_i$, average peak intensity calculated from three replicates analysed for each sample to ensure reproducibility) by the total intensity of all peaks ($\Sigma\bar{h}$) in the electropherogram (Equation 1).

$$I_i = \frac{\bar{h}_i}{\Sigma\bar{h}} \times 100\% \quad (1)$$

The second filtering criterion was that the relative peak ratio had to exceed 1% to be retained for further analyses.

PRIMER 6 statistical software (PRIMER-E Ltd., Plymouth Marine Laboratory, Plymouth, U.K.) was used to perform multivariate statistical analysis. Pairwise Bray–Curtis's similarity scores between LH-PCR profiles were calculated based on the square root transformed data (Clarke et al. 2006). The similarity score is conventionally defined to take values in the range from 0 (if two samples are totally dissimilar) to

100% (if two samples are totally similar). The resulting Bray-Curtis similarity matrices were then used for hierarchical agglomerative cluster analysis (CLUSTER) with the results displayed as group average dendrograms (Jain et al. 1999). Non-metric multidimensional scaling (NMDS) based on Bray-Curtis similarity scores was performed as an unconstrained ordination method to graphically visualise inter-sample relationships (Clarke 1993). Adequacy and accuracy of NMDS representation of samples relationship can be assessed by the Stress value. The Stress value < 0.05 gives an excellent representation with no prospect of misinterpretation. The Stress value > 0.3 indicates that the points were arbitrarily placed in the 2-dimentional ordination space. The examining differences between *a priori* defined groups was performed by 'Analysis of Similarities' (ANOSIM) on data based on factors of soil collection site, applied extraction kit, storage length and temperature, and treatment of soil samples with IPA (Clarke & Green 1988). The ANOSIM test statistic R value usually falls between 0 and 1, indicating the significance of difference. For example, if all replicates within a site are more similar to each other than any replicates from different sites then R=1; opposite, R equals approximately zero if the intra- and intergroup similarities of the samples' profiles are of the same level (similar) in average, this means that the null hypothesis of no difference between groups is true.

## 2.3 Results and discussion

### 2.3.1 Selection of DNA polymerase with the lowest amount of residual bacterial DNA

The extreme sensitivity associated with PCR can lead to problems of false positive amplification results caused by inadvertent exogenous contamination with analysis unrelated DNA (Champlot et al. 2010).

Sources of DNA contamination can be different and hard to predict. Inappropriate handling of samples of interest is one of the possible reasons for exogenous DNA contamination at the very first stages of the analysis. Such type of contamination is prevented by wearing gloves or whole-body suits, which is routine for sampling at crime scenes. Contamination of laboratory surfaces, instrumentation and plastic-ware, is also a known problem and therefore such decontamination procedures as the use of UV-irradiation and bleach treatment have to be implemented on a regular basis in any PCR laboratories (van Oorschot et al. 2010; Ballantyne et al. 2013). PCR products carry-over is the most common type of contamination and is simple to prevent by organising a strict separation of pre- and post-amplification areas and equipment. Contamination of PCR reagents and DNA extraction kits with bacterial DNA traces is a major problem when broad-range or unspecific primers are used (Klaschik et al. 2002; Iulia & Bianca 2013). Many publications to date report that commercial DNA polymerase preparations inevitably contain trace amount of contaminating microbial DNA (Mühl et al. 2010; Heo & Kim 2013; Takahashi et al. 2014). Since universal bacterial primers are most commonly used for soil microbial community DNA typing techniques, for example 16S rRNA sequencing and LH-PCR based DNA profiling, it would be highly relevant to estimate the presence of bacterial DNA traces in available

DNA polymerases preparations and to select the one with the lowest amount of residual bacterial DNA present.

Eight available DNA polymerases were assessed (Table 2.1). The assessment was performed using two different sets of universals 16S rRNA gene-specific primers. This was due to the fact that different primer combinations exhibit different performance in regard to reliable identification of trace amount of bacterial DNA (Iulia & Bianca 2013). The following primer pairs were tested: pair #1 27-F and 355-R and pair #2 cc-F and cd-R (see Table 2.2). These pairs were designed for amplification of the V1-V2 and V3-V5 hypervariable domains of 16S rRNA bacterial gene, respectively. Figure 2.2 and Figure 2.3 show the results of PCR amplification of ten negative control samples along with two positive control samples with 27-F and 355-R, and cc-F and cd-R primers, respectively. A summary of the testing for all available DNA polymerases in combination with two outlined above primer sets is presented in Table 2.4.

Based on the data obtained, the most 'DNA-free' DNA polymerases appeared to be the HotStar Taq DNA polymerase from Qiagen, Q5 High Fidelity DNA polymerase from NEB and GoTaq DNA polymerase from Promega. None of those gave false positive amplification signals in the no-template control amplification reactions with either of primer pairs 27-F/355-R or cc-F/cd-R (Figure 2.2A, C, E, F, and Figure 2.3A, B, C, D, F, G respectively). The HotStar Taq DNA polymerase from Qiagen was chosen for use in subsequent DNA typing of soil microbial community.

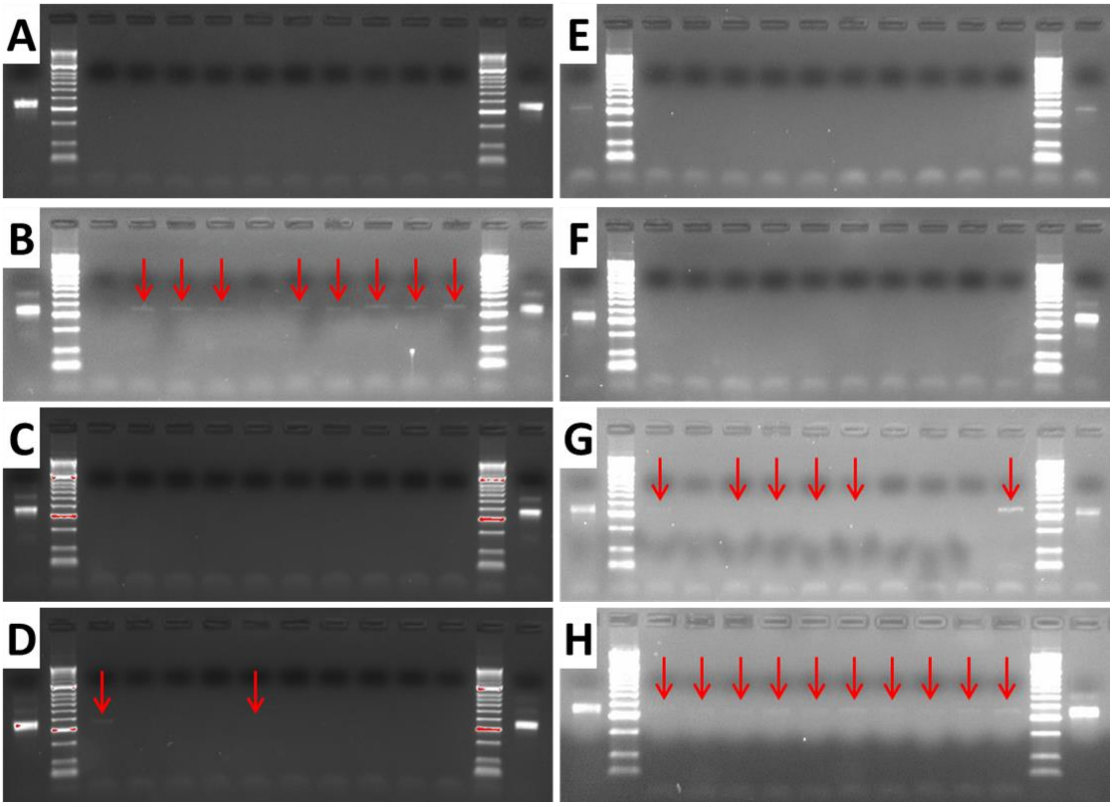



**Figure 2.2. Results of the assessment of eight available DNA polymerases for the presence of bacterial DNA traces using 27-F and 355-R primers.** (A) HotStar Taq DNA polymerase, Qiagen; (B) AmpliTaq Gold DNA polymerase, Life technologies; (C) Q5 DNA polymerase, NEB; (D) MyTaq DNA polymerase, Bioline; (E) OneTaq DNA polymerase, NEB; (F) GoTaq DNA polymerase, Promega; (G) Mango Taq DNA polymerase, Bioline; (H) Taq DNA polymerase, NEB. Each electropherogram (2% agarose, stained with EtBr) contains from left to right: positive amplification control, DNA ladder (Hyper Ladder II, Bioline), 10 no-template controls, DNA ladder (Hyper Ladder II, Bioline), positive amplification control. Red arrows show the presence of the specific amplification product (could be invisible in a hard copy version of the thesis) of the same size as in the positive amplification control.

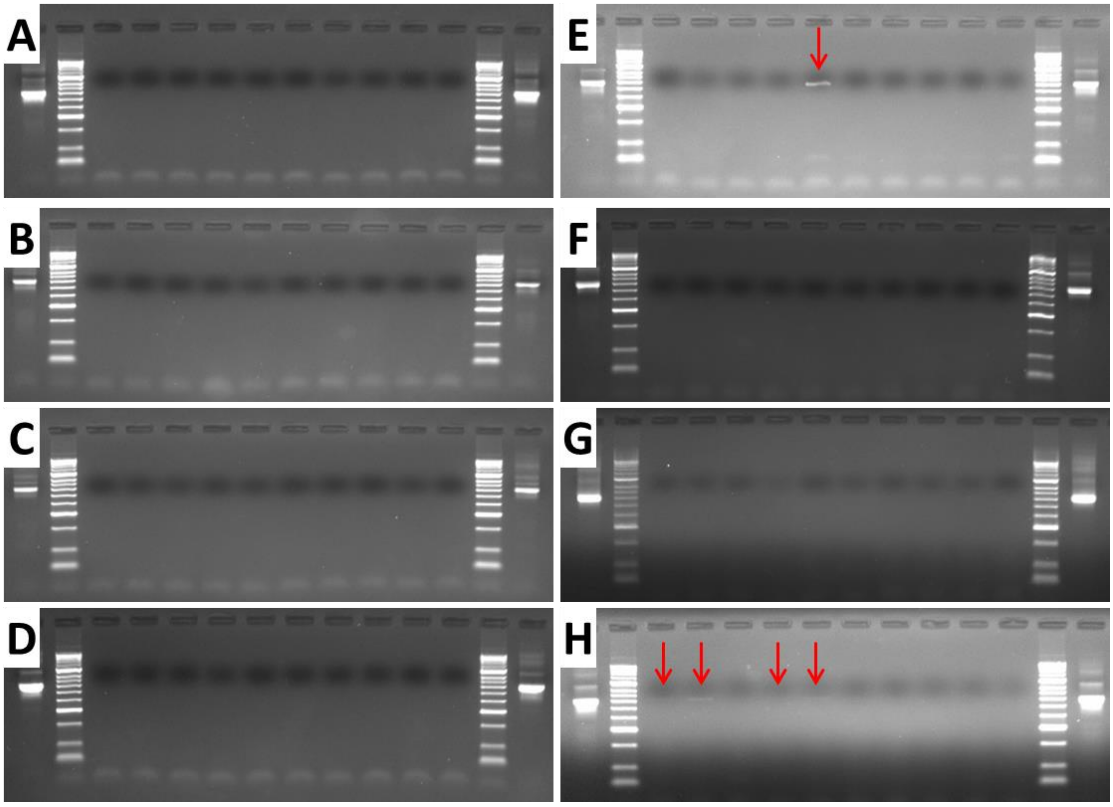



**Figure 2.3. Results of the assessment of eight available DNA polymerases for the presence of bacterial DNA traces using cc-F and cd-R primers.** (A) HotStar Taq DNA polymerase, Qiagen; (B) AmpliTaq Gold DNA polymerase, Life technologies; (C) Q5 DNA polymerase, NEB; (D) MyTaq DNA polymerase, Bioline; (E) OneTaq DNA polymerase, NEB; (F) GoTaq DNA polymerase, Promega; (G) Mango Taq DNA polymerase, Bioline; (H) Taq DNA polymerase, NEB. Each electropherogram (2% agarose, stained with EtBr) contains from left to right: positive amplification control, DNA ladder (Hyper Ladder II, Bioline), 10 no-template controls, DNA ladder (Hyper Ladder II, Bioline), positive amplification control. Red arrows show the presence of the specific amplification product (could be invisible in a hard copy version of the thesis) of the same size as in the positive amplification control.

**Table 2.4. Number of false-positive amplification results obtained after amplification of ten no-template controls with corresponding DNA polymerase and a primer pair.**

| DNA polymerase | Primer pair | |
|---|---|---|
| | ***cc-F and cd-R*** | ***27-F and 335-R*** |
| HotStar Taq, Qiagen | **0/10** | **0/10** |
| Q5, NEB | **0/10** | **0/10** |
| MyTaq, Bioline | **2/10** | **0/10** |
| OneTaq, NEB | **0/10** | **1/10** |
| GoTaq, Promega | **0/10** | **0/10** |
| Mango, Bioline | **6/10** | **0/10** |
| Taq, NEB | **10/10** | **4/10** |
| AmpliTaq Gold 360, Life Technologies | **8/10** | **0/10** |

### 2.3.2 Selection of an effective soil DNA extraction method for the assessment of microbial community.

Total DNA was extracted from the five replicates of each soil type by five commercial DNA extraction kits, namely PowerSoil **(PS)** DNA Isolation kit, UltraClean **(UC)** Soil DNA Isolation kit , EZNA **(EZ)** Soil DNA kit, ZR Soil Microbe DNA kit (**ZR**) and FavorPrep **(FP)** Soil DNA Isolation kit. Altogether 75 different DNA extracts were obtained.

Co-extraction of contaminants such as humic compounds (Figure 2.4) is a major problem associated with various soil metagenomics studies. These humic substances accumulate in soil as a result of plant, animal and microbial biomass decomposition. Depending on the soil type and the DNA extraction kit employed the amount of these contaminants may vary significantly in the DNA extracts, being even sometimes evident by brown colour.

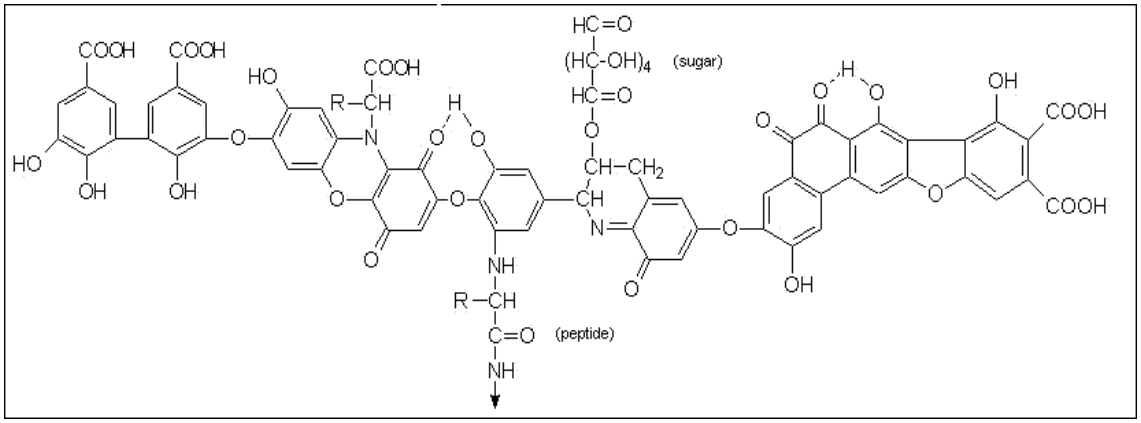



**Figure 2.4. Model structure of humic acid (**reproduced from Stevenson 1982)

Thus, in this study, almost all DNA extracts obtained by the FP kit had brown colour, clearly indicating the presence of humic contaminants in high concentration. These DNA preparations, as well as the extraction kit, were eliminated from subsequent analysis. This decision was motivated by the well-known fact that the presence of the humic compounds has a great impact on the subsequent both physical-chemical and

enzymatic investigation (Jackson et al. 1997; Dong et al. 2006; Miao et al. 2014). For instance, the humic substances can interfere with the quantification of DNA (Rajendhran & Gunasekaran 2008). This is likely explained by multiple aromatic rings present in the structure of humic compounds which could contribute to the overall absorbance of the DNA extracts in a range of 240 - 340 nm resulting in unreliable DNA quantification using UV-VIS spectrometry.

It has been reported that as little as 1 ng of humic substances per reaction can inhibit PCR amplification (Menking et al. 1999). This most likely happens due to chelation of $Mg^{2+}$ ions by multiple carbocyclic functionalities in humic acids (Alaeddini 2012).

DNA purity of the extracted samples was evaluated by measurement of the absorbance ratios: A260/280 and A260/230 (Table 2.5). The presence of humic acids was estimated by measuring the absorbance of the DNA preparations at 320 and 340 nm (A320 and A340, respectively). Comparison of the purity indices, as well as DNA yield, for different extraction kits, showed large differences depending on the soil type and the extraction kit used. It was observed that different amounts of DNA, normalized to 1 g of soil, were extracted from different soil types. Thus, for the soils A and B, rich with organic material the DNA yield was higher, than for the soil C. The highest DNA yield for soils A and B was obtained for PS (9.9 ± 1.9 µg/g soil and 7.9 ± 1.7 µg/g soil, respectively) and ZR (7.2 ± 2.0 µg/g soil and 4.4 ± 2.0 µg/g soil, respectively) extracts (Table 2.5) DNA yield of UC and EZ extracts for these soils was in average below 2 µg/g soil. These data clearly indicate the differences between the different soil DNA extraction kits in their ability to recover DNA from soil. This difference between the kits performances was less significant for the soil C, which is likely explained by the low amount of organic material present in this soil type.

**Table 2.5. DNA extraction efficiency of the commercial soil DNA extraction kits.**

| Soil DNA Extraction Kit | Soil sample | A260/A280 | A260/A230 | A320 | A340 | DNA yield, µg/g soil | Positive PCR amplification obtained from |
|---|---|---|---|---|---|---|---|
| PowerSoil (PS) | A | 1.60 ± 0.08 | 1.28 ± 0.24 | 0.10 ± 0.05 | 0.07 ± 0.06 | 9.9 ± 1.9 | 2-fold dilution |
| | B | 1.67 ± 0.10 | 1.27 ± 0.24 | 0.04 ± 0.03 | 0.03 ± 0.02 | 7.9 ± 1.7 | no dilution required |
| | C | 1.29 ± 0.05 | 0.58 ± 0.18 | 0.05 ± 0.01 | 0.04 ± 0.01 | 2.9 ± 0.6 | no dilution required |
| UltraClean (UC) | A | 1.25 ± 0.04 | 0.75 ± 0.04 | 0.24 ± 0.06 | 0.19 ± 0.05 | 2.8 ± 1.0 | 10-fold dilution |
| | B | 1.23 ± 0.18 | 0.45 ± 0.13 | 0.08 ± 0.03 | 0.07 ± 0.02 | 1.9 ± 0.8 | no dilution required |
| | C | 1.22 ± 0.03 | 0.64 ± 0.10 | 0.21 ± 0.06 | 0.16 ± 0.04 | 1.1 ± 0.1 | 10-fold dilution |
| ZRSoil (ZR) | A | 1.54 ± 0.15 | 0.72 ± 0.37 | 0.15 ± 0.07 | 0.09 ± 0.04 | 7.2 ± 2.0 | no dilution required |
| | B | 1.63 ± 0.21 | 0.52 ± 0.26 | 0.07 ± 0.02 | 0.04 ± 0.01 | 4.4 ± 2.0 | no dilution required |
| | C | 2.00 ± 0.33 | 0.40 ± 0.13 | 0.03 ± 0.01 | 0.03 ± 0.01 | 2.0 ± 0.9 | no dilution required |
| EZNASoil (EZ) | A | 1.32 ± 0.08 | 0.60 ± 0.28 | 0.23 ± 0.12 | 0.19 ± 0.10 | 2.0 ± 0.5 | 10-fold dilution |
| | B | 1.39 ± 0.08 | 0.57 ± 0.35 | 0.19 ± 0.06 | 0.15 ± 0.05 | 1.8 ± 0.5 | 10-fold dilution |
| | C | 1.67 ± 0.35 | 0.28 ± 0.16 | 0.04 ± 0.03 | 0.04 ± 0.02 | 1.0 ± 0.1 | 2-fold dilution |

Data represented as Average ± Standard Deviation (SD) for five replicate extracts of each soil type obtained by each DNA extraction kits.

Having in mind that the DNA extracts will be used further in different enzymatic assays, it was decided to perform qualitative estimation of inhibitors content via PCR amplification. As such, PCR amplification of the DNA preparations obtained by four extraction kits was performed. Each initial DNA extract and its two-fold and ten-fold dilutions were amplified with cc-F and cd-R primers (Table 2.2) and Qiagen HotStar Taq DNA polymerase, chosen after DNA polymerase evaluation experiment. Little correlation between the PCR results and the obtained UV-VIS indices was observed. For example, amplification of undiluted DNA extracts from samples B and C obtained

with PS and ZR kits showed positive results (Figure 2.5 and Figure 2.6, respectively). In this case for both kits absorbance at 320 and 340 nm was less than 0.07 and 0.04, respectively.

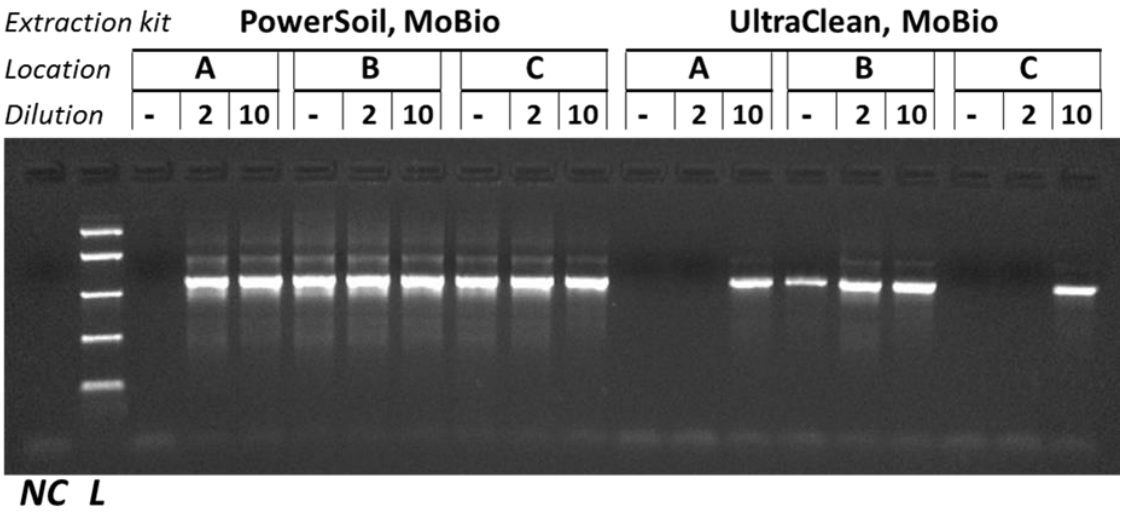



**Figure 2.5. Electrophoretic picture showing results of PCR amplification of soil DNA extracts obtained using PowerSoil (MoBio, USA) and Ultraclean (Mobio, USA) DNA extraction kits. NC = negative control; L = DNA ladder (Easy Ladder II, Bioline); No dilution, two- and ten-fold dilutions of the extracted DNA specimens designates as "-", "2" and "10", respectively, were used for PCR amplification.**

Amplification of the undiluted DNA preparations of soil A extracted with these kits showed partial, for the ZR kit, and complete, for the PS kit, PCR inhibition. A320 and A340 values for the ZR extracts of soil A were at least as twice as for soils B and C (0.15 and 0.09, respectively). Total DNA from the soil A extracted with PS kit was characterised by relatively high A260/A280 and A260/A230 ratios (1.6 and 1.28, respectively) along with low A320 and A340 absorbances of 0.1 and 0.07, respectively, but still led to PCR inhibition. This inhibition effect was likely due to the high DNA yield of the extracts and therefore relatively high amount of co-extracted humic acids presence which was sufficient to inhibit PCR. A weak band of the appropriate size was identified after the amplification of undiluted soil B extracted with the UC kit (A320 = 0.08, A340 = 0.07), its two times dilution showed better PCR performance (Figure 2.5).

Interestingly, for soil C extracted by EZ kit A320 and A340 values were measured to be 0.04 and 0.04, respectively. However, despite of these low values the PCR amplification of this undiluted sample failed. This is likely to be explained by the value of A260/230 ratio of 0.28 that was the lowest among all assessed DNA extracts, which can also serve as an indicator of inhibitors presence. To achieve positive amplification results the samples obtained with the UC or EZ kits required two-fold and ten-fold dilutions to remove PCR inhibition and were characterised by levels of A320 and A340 absorbance more than 0.19 and 0.15, respectively (Figure 2.5 and Figure 2.6).

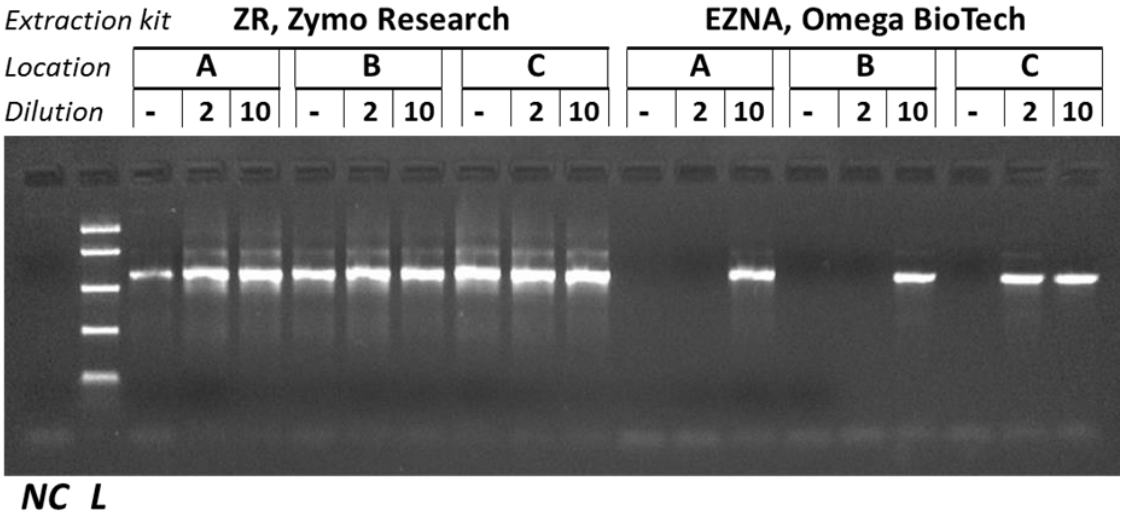



**Figure 2.6. Electrophoretic picture showing results of PCR amplification of soil DNA extracts obtained using ZR (Zymo Research, USA) and E.Z.N.A (Omega, USA) extraction kits. NC = negative control; L = DNA ladder (Easy Ladder II, Bioline); No dilution, two- and ten-fold dilutions of the extracted DNA specimens designates as "-", "2" and "10", respectively, were used for PCR amplification.**

PCR amplification of extraction blank controls (EBC) that were processed along with the soil samples extraction gave negative results for PS, UC and ZR kits (Figure 2.7). However, PCR products were detected in EBC of EZ kit. This might indicate that some background DNA was introduced during the extraction process. This can be likely explained by that the DNA extraction procedure for this kit involved weighing glass beads, 2-propanol DNA precipitation followed by drying on air and the use of user-

provided DNA collection test-tubes. Any of those creates the possibility for exogenous DNA to be introduced.

The data obtained suggest that the DNA yield, as well the quality and purity of the soil DNA preparations, depend on both compositions of the soil sample and the efficacy of the extraction kit employed. Based on the obtained DNA yields, purity indices and performance in PCR amplification, two soil DNA extraction kits, namely PowerSoil from MoBio and ZR Soil Microbe from Zymo Research, were shown to be successful at extracting high-purity DNA with a good yield. Soil DNA extracts obtained by these two kits were further evaluated with LH-PCR technique.

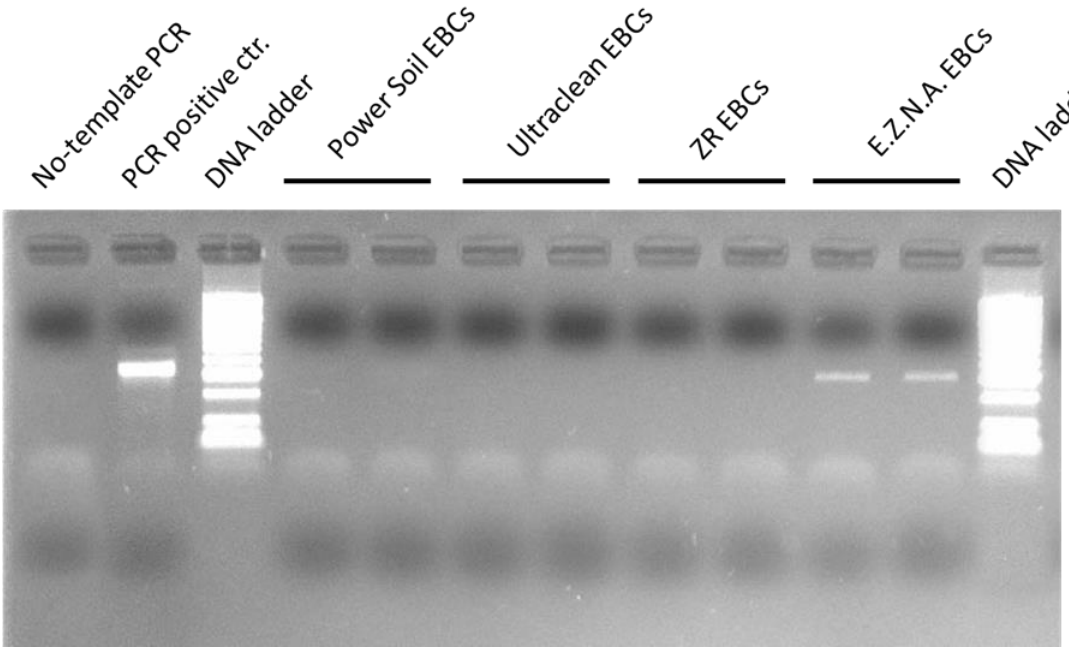



**Figure 2.7. Extraction blank controls (EBC) amplification results.** Lane 1 – no-template amplification control, lane 2 – positive amplification control, lane 3 – DNA ladder (HyperLadder II, Bioline), lanes 4-5 EBCs PowerSoil DNA extraction kit, lanes 6-7 - EBCs Ultraclean DNA extraction kit, lanes 8-9 - EBCs ZR DNA extraction kit, lanes 10-11 –EBCs E.Z.N.A DNA extraction kit, lane 12 – DNA ladder (HyperLadder II, Bioline).

### 2.3.3 LH-PCR profiling of soil metagenomic DNA extracted by PowerSoil and ZR soil microbe DNA extraction kits.

There is evidence that different soil DNA extraction methods may recover different microbial populations of soil community and that DNA originating from different species might be released differently (Delmont et al. 2011).

Length heterogeneity PCR (LH-PCR) profiling method was found to be fast, robust and reproducible method for the analysis of soil microbial diversity and was previously tested for forensic application (Moreno et al. 2011). In this thesis, this approach was used to assess the influence of soil DNA extraction method on obtaining representative and stable DNA preparations suitable for further analysis using modern metagenomic HTS-based approaches. The LH-PCR method is based on the analysis of natural variations in the length of gene regions belonging to specific domains within genome. For example, analysis of the 16S rRNA gene is the most commonly used method for analysis of bacterial and archaeal communities (Pace 1997). In this study, the LH-PCR amplification was performed with 27-F and 355-R primers that target hyper-variable V1 and V2 regions within bacterial 16S rRNA gene (Moreno et al. 2006). The forward primer was labelled at its 5' terminus with the FAM fluorophore (Table 2.2). This allowed for analysis of the obtained PCR amplification products on a 3130*xl* Genetic Analyser. A comparison of the profiles was based on the number and intensities of peaks detected in the resulting electropherograms; these peaks corresponded to the PCR products of different length.

Typical soil microbial community LH-PCR profiles for soil samples from locations A, B and C are illustrated in Figure 2.8. The length of LH-PCR fragments varied from 311 to 353 bp were highly reproducible, and resulted in 23 reproducibly observed peaks obtained across all soil types (Appendix A, Tables A1, A2). The

number of peaks in LH-PCR soil profiles obtained in current study were similar to the studies where the same primers for the LH-PCR profiling of soil were used (Chaudhary, 2012; Ritchie 2000).

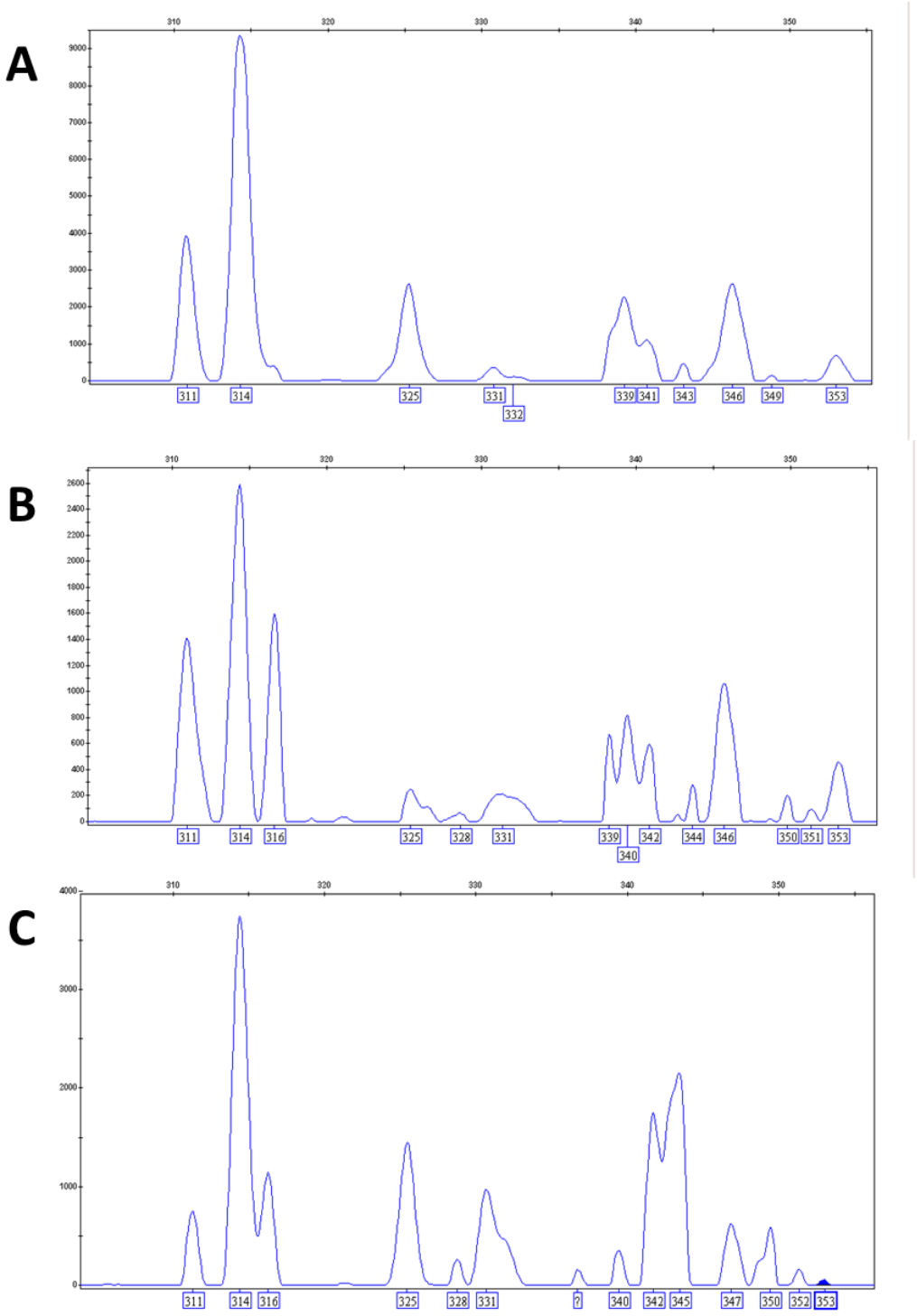



**Figure 2.8. Typical LH-PCR profiles from soil samples taken from locations A, B and C**. The x axis is in bps and the y axis is RFU.

In order to perform a statistical comparison of the obtained LH-PCR profiles, Bray-Curtis pair-wise similarity scores were calculated. The test's performance is based on the fact that identical samples have 100% similarity and samples with no common variables (or peaks found in LH-PCR profiles in this study) have 0% similarity (Clarke & Warwick 2001). The CLUSTER analysis demonstrated clear separation of samples into clusters according to the soil sampling site and extraction kits used (Figure 2.9A). Every profile on NMDS ordination plot is represented by a point while relative distances between the points reflect the relative dissimilarities between these particular profiles (Clarke & Warwick 2001). On the NMDS plot (Figure 2.9B) three large clearly defined clusters corresponded to three soil sampling sits are clearly observed. Within each of these clusters two sub-clusters were seen, which were associated with the extraction kit applied.

LH-PCR profiles from soils B and C replicative extracts obtained by each of PS and ZR kits were very similar (95%) (Figure 2.9A). This result clearly indicates the reproducibility of both DNA extraction and PCR amplification procedures. An average similarity of 85% and 93% was observed between the replicative extracts from the soil A obtained by ZR and PS kits, respectively. Whereas approximately 31% dissimilarity was found between the LH-PCR profiles from ZR and PS extracts of soil A. The same comparison was performed for the samples B and C which showed dissimilarity between the kits of 13% and 19%, respectively.

The results obtained support the previous findings that different extraction methods isolate different populations of the soil microbial community (Delmont et al. 2011). At the same time, distances seen on the NMDS plot between the profiles from the same soil samples obtained with different extraction methods were smaller than the distances between the profiles form different soils types. Thus, differences between the

LH-PCR profiles from three soil types regardless of the extraction kit used was shown statistically significant by ANOSIM analysis (Global R=0.991, p<0.0001), indicating that the provenance of soil samples is more important at differentiation of the soils then the DNA extraction kit applied.

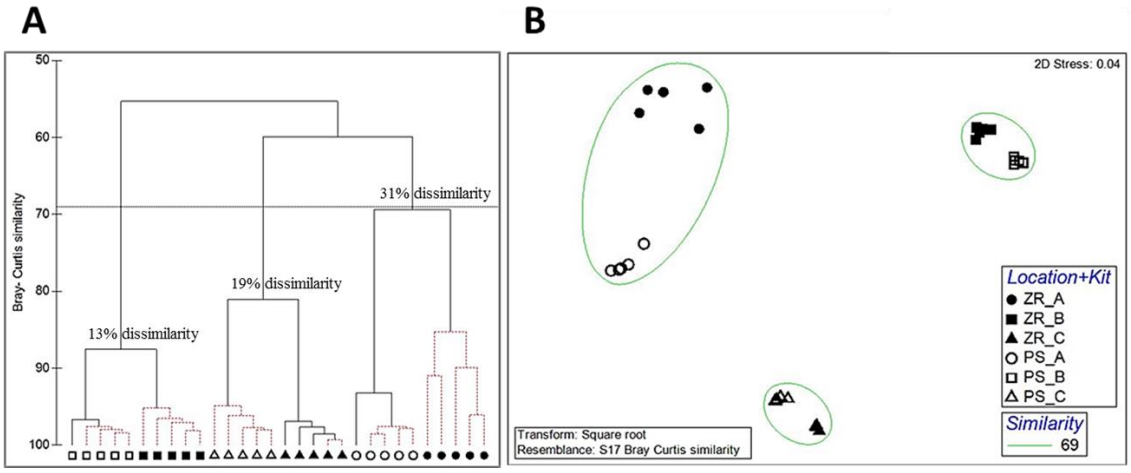



**Figure 2.9. Cluster analysis (A) and non-metric multidimensional scaling (NMDS) (B) plot** of LH-PCR profiles of total DNA extracted from the soil samples A (circles), B (squares) and C (triangles) by PowerSoil (hollow symbols) and ZR soil microbe (solid symbols) DNA extraction kits.

High similarity of the LH-PCR profiles obtained by different extraction kits from the same soil samples and their correct clustering according to the sampling locations showed that these extraction kits performed similarly in terms of extraction of the most representative DNA from the soil bacterial community. The PS extraction kit is most widely used method for extracting DNA from soil samples (Nemergut et al. 2011; Aanderud et al. 2013; Mason et al. 2014), however based on the results obtained in the current study, a recommendation is that both PS and ZR extraction kits are suitable methods for the extraction of high-quality total microbial DNA from soil. ZR soil Microbe DNA extraction kit was chosen for subsequent studies in this thesis.

### 2.3.4 Evaluation of the effect of storage conditions on the soil microbial community DNA typing.

The impact of soil pre-treatment and storage conditions on soil DNA typing were investigated on three types of soil samples taken from distinct locations. This included washing of soil samples with 2-propanol (IPA), different temperatures and length of the following storage.

Aliquots of each soil type with and without IPA washing were kept at three different temperatures, namely 25 $^{o}$C (room temperature, RT), 4 $^{o}$C and -20 $^{o}$C, for two and four weeks before DNA extraction. LH-PCR profiling of the obtained DNA extracts was then used for the monitoring of changes in soil microbial structure (performed in triplicates).

Evaluation of the storage effect was based on the statistical analysis of differences between the LH-PCR profiles obtained for fresh and stored samples. CLUSTER analysis based on Bray-Curtis profiles similarity scores demonstrated correct separation of all LH-PCR profiles into three clusters according to the soil collection sites (Figure 2.10). Within these clusters, the samples underwent DNA extraction within 24 hrs from collection were randomly mixed with those kept at different temperatures for different time periods.

Storage of soil samples under different conditions resulted in a little alteration of the identified microbial composition compared with the fresh soil DNA extracts, as it also seen on the NMDS plots (Figure 2.11).

The LH-PCR profiles generated from the aliquots were found to be of 85% similar, regardless of whether they were washed with IPA (Figure 2.11A), length of storage (Figure 2.11B) and temperature (Figure 2.11C). These was true for all soil types, except two outliers corresponding to the profiles from soil samples A stored at

RT and +4 °C for two weeks as seen on NMDS plots. However, the same soil type samples stored at the same conditions for four weeks were quite similar to the initial samples, therefore the outliers could likely be explained as PCR artefacts.

To test the hypothesis about difference between the LH-PCR profiles based on soil collection site, storage length and temperature, and soil samples treatment with IPA was performed by ANOSIM analysis. Again the LH-PCR profiles from different soil types were confirmed being significantly different (Global R=0.997, p <0.001), whereas the effect of the storage factors was found insignificant (R values were close to 0 and p > 0.05).

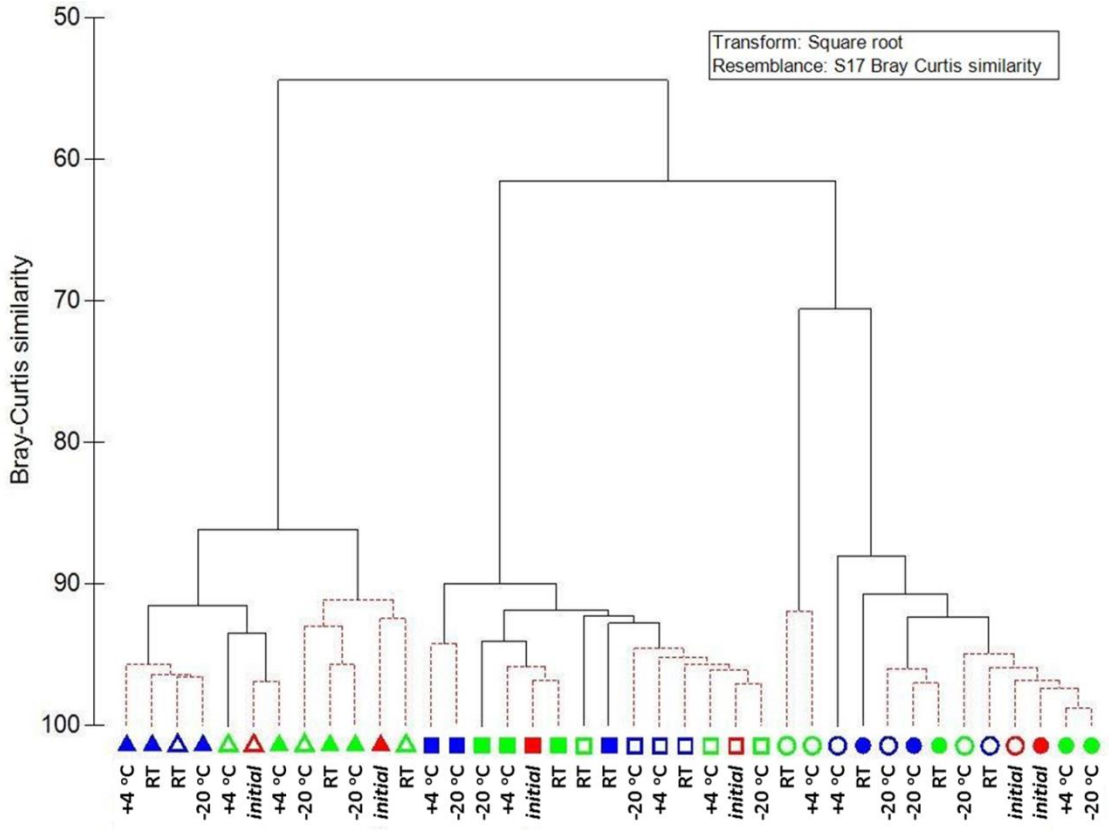



**Figure 2.10. CLUSTER analysis of LH-PCR profiles obtained from soil samples stored at different conditions.** Circles (○, ●), squares (□, ■) and triangles (Δ,▲) denote soil samples A, B and C, respectively. Solid and hollow symbols denote soil samples treated (solid) and non-treated (hollow) with 2-propanol. Symbols encoded in colour denote samples used in the DNA extraction within 24 hrs after collection (red symbols), or stored for 2 weeks (green symbols) and 4 weeks (blue symbols) at the room temperature (RT), +4 $^{o}$C or -20 $^{o}$C (as shown on the dendrogram).

In this study three different microbial communities extracted from soils stored under different storage conditions were assessed. It was observed that the whole bacterial community structure had not changed significantly during the storage in comparison with the structure obtained from the ‘fresh’ samples. Our results, therefore, support other findings indicating that the origin of samples has a greater effect on the soil metagenomic composition and does not significantly depend on the conditions under which samples are stored prior to DNA extraction (Lauber et al. 2010; Rubin et al. 2013).

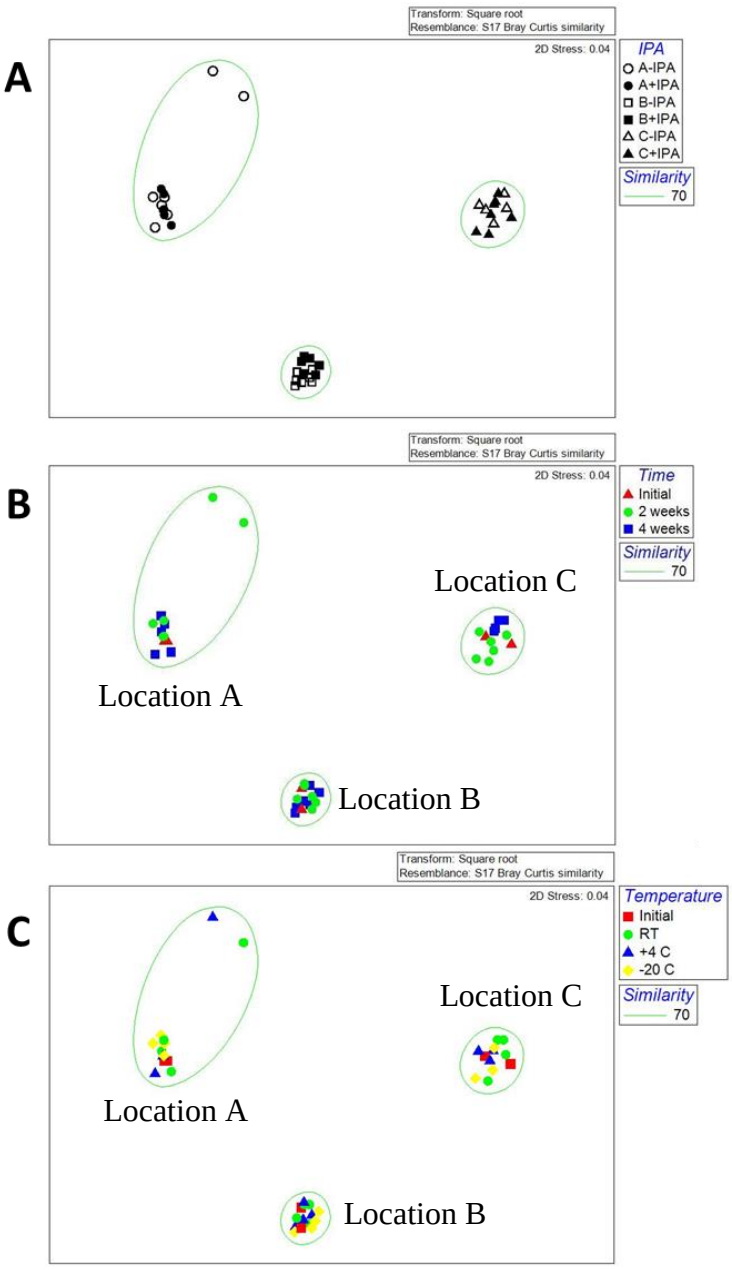



**Figure 2.11. Non-metric multidimensional scaling (NMDS) plots of LH-PCR profiles of total DNA extracted form soil samples A, B and C and stored at different conditions.** NMDS plots are separated by different factors: (A) treatment of soil samples with 2-propanol (IPA), solid and hollow symbols denote LH-PCR profiles generated from total DNA extracted from the soil samples treated and non-treated with IPA, respectively; (B) soil samples storage duration, where red, green and blue symbols denote LH-PCR profiles generated from total DNA extracted from the soil samples within 24 hours after collection, two and four weeks of storage, respectively; (C) soil samples storage temperature, where red symbols denote LH-PCR profiles generated from total DNA extracted from the soils samples within 24 hrs after collection, while green, blue and yellow ones represent LH-PCR profiles generated from total DNA extracted from the samples stored at room temperature, +4 $^{o}$C and -20 $^{o}$C, respectively.

## 2.4 Conclusions

The study presented in this chapter was aimed to assess the impact of biases that can be introduced in soil DNA typing at the stage of soil sample handling. This includes storage of the collected samples and DNA extraction. There are a few controversial reports on the influence of these processes on interpretation of results of soil community DNA analysis (Feinstein et al. 2009b; Lauber et al. 2010; Rubin et al. 2013). The sources of these variations must be considered and evaluated in order to soil discrimination methods based on DNA profiling could become a useful forensic tool and pass admissibility test to be accepted in court.

Results presented here clearly show that selection of the DNA extraction kit is of paramount importance for obtaining reliable and reproducible picture of soil microbial community composition. Out of five different soil DNA extraction kits only two, namely PowerSoil and ZR soil microbe DNA extraction kits, were successful at providing DNA of high yield and quality suitable for subsequent applications. The similar results were obtained for three different soil types.

Bacterial DNA profiles of the extracted DNA preparations were obtained using LH-PCR techniques and then were shown to be consistent across two extraction kits applied for the same soil type. In addition, replicate DNA extractions for each of the extraction kits resulted in highly reproducible LH-PCR profiles. The assessment of sample storage conditions, such as temperature, storage length and washing the soil sample with 2-propanol prior DNA extraction also did not reveal significant influence on the soil DNA LH-PCR fingerprints. It was shown that LH-PCR profiles of soil subsamples even stored at room temperature for four weeks were more similar to those generated from freshly collected soil samples rather than to the soils originated from other locations.

The LH-PCR technique appeared to be a very useful method for rapid evaluation of soil microbial composition. The main limitation of LH-PCR profiling is that the DNA of only dominant members of the community is assessed, but the majority of low–abundant members of soil microbiome might be undervalued. Fine-scale community resolution is therefore needed for better understanding distribution and spatial and temporal variations of rare taxa. In addition, the inherent drawback of the method is that it is unknown whether two peaks with the same retention time derived from two different electropherograms have the same sequence and match by chance. Moreover, a single peak observed on the electropherogram might have a number of PCR products of the same length from different unrelated species persisting in such a complex DNA source as soil. Therefore the judgment about the identity of the samples based only on the comparison of the PCR product lengths will be potentially compromised and due to potential false positive, will most likely not be accepted in the court.

In conclusion, forensic soil DNA analysis requires standardised protocols for handling of soil samples and soil DNA extraction to ensure robust, reliable and reproducible results to be obtained. At the level of the given research, it is evident that different commercially available soil DNA extraction and amplification reagents perform within a wide range of variability. In order to secure obtaining reliable results of soil microbial community analysis, the issues related to the background microbial contamination need to be addressed.

The presented small-scale rational selection of the proper reagents could serve as a model system for a further broad-scale evaluation based on the more reliable and trusted method, such as one of the high throughput DNA sequencing approaches to be selected, aimed to the development of a standardised procedure of soil evidential samples processing.

# Chapter 3. 16S rRNA sequencing for forensic soil discrimination

## 3.1 Introduction

Targeted metagenomics is the most commonly used approach for the investigation of soil microbial community structure using modern high throughput DNA sequencing (HTS) techniques (Suenaga 2012). The method involves the PCR amplification of highly conserved genes and genomic regions. Ribosomal RNA genes such as the small subunit (SSU) and large subunit (LSU) are the most widely used markers for taxonomic analysis of microbial species. In order to specifically cover different taxonomic groups, loci such as the 16S rRNA gene in bacteria and archaea, the ITS in fungi, the tRNA-Leu gene in plant genomes are targeted (Epp et al. 2012).

The 16S rRNA gene is highly conserved among all bacteria and archaea. The length of the gene varies in different bacterial species and consists on average of 1500 bp. It is characterised by the presence of multiple distinct hypervariable domains (V1-V9) that makes this locus perfectly suitable for HTS and bioinformatics analysis (Petrosino et al. 2009; Tringe & Hugenholtz 2008).

Bioinformatic analysis of 16S sequencing data can be performed in two different ways (Chen et al. 2013). The first relies on direct annotation of the obtained sequencing reads against reference rRNA databases containing known sequences from previously characterised microorganisms. Among the most commonly used databases are Greengenes (DeSantis et al. 2006), RDP (Cole et al. 2014) and SILVA (Quast et al. 2013). All of these are publically available and contain both bacterial 16S (Greengenes, RDP, SILVA) and eukaryotic 18S (SILVA) reference sequences. The application of this approach for the analyses of a highly diverse soil microbial community, where 99% of microorganisms are unknown and uncultivable, is limited by the incompleteness of the existing reference databases.

Another approach based on the determination of operational taxonomic units (OTU) has also gained a wide acceptance in ecological research (Schloss & Westcott 2011). In this method all sequences are first grouped into unique OTU clusters at the user defined similarity level. For example, it is widely accepted that OTU clusters with sequences similarly of 97% and higher correspond to taxonomic units at the species level. The longest representative sequence is then selected for each cluster, which essentially represents an OTU. Further, the resultant sets of representative OTUs can be used for statistical analysis, characterisation of diversity and evaluation of the richness of the microbial community. A distinctive feature of the OTU-based method is that all sequencing reads need to be assigned into OTU clusters. This makes the OTU-based method highly useful for the analysis of less characterised and complex metagenomes, for example soil metagenome. Moreover, in order to find out taxonomic composition of the microorganisms present in the metagenome, the representative OTUs can be subsequently annotated using the same reference rRNA databases as described above.

Quantitative Insights Into Microbial Ecology (QIIME) is an open-source software pipeline (http://qiime.sourceforge.net/), running under the Linux operational system that allows for a wide range of manipulations and analyses of amplicon-based metagenomic sequencing data. Bioinformatics analysis of the data in QIIME starts from the raw or pre-processed reads. Pre-processing usually involves adapter and primer trimming and filtering of the reads containing ambiguous and low-quality bases. A Phred quality score of 20 (Q20), assigned to a base by the sequencing platform, equates to a base call accuracy of 99% (Ewing et al. 1998; Ewing & Green 1998; Cock et al. 2010). Q20 is an acceptable score for sequencing data and commonly used by sequencing facilities.

As with most emerging technologies, targeted metagenomics is being taken from ecological research and adapted by forensic community for its particular needs (Budowle et al. 2014). Researchers from Italy investigated the potential of HTS for forensic identification of soils along with traditional soil characterisation such as colour, polarised microscopy and X-ray diffraction (Giampaoli et al. 2014). The authors investigated eukaryotic and bacterial communities and showed that bacterial markers allowed for discrimination of geologically similar soils from distinct environments. An Estonian research group evaluated two eukaryotic markers (18S rRNA gene V2-V3 region and SSU rRNA region for arbuscular mycorrhizal (AMF) for fungi) and a bacterial marker (V2-V3 region for 16S rRNA gene) for the analysis of microbial communities in soil samples taken from different environments namely forests, fields, grasslands and a town park (Lilje et al. 2013). They showed that the 18S eukaryotic marker was more efficient and flexible than the 16S bacterial marker and AMF fungal marker. The other strong point of the study was multi-replicate sampling of each area.

Young *et al*. tested the ability of four molecular markers, bacterial 16S rRNA, eukaryotic 18S rRNA, plant trnL, and fungal ITS1, to distinguish two contrasting soil types taken 14 km apart (Young et al. 2014). One soil represented a dark, organic-rich, saline mangrove soil whereas the other one was from a coastal sandy site with low water, nutrient and organic content. From each site the authors collected four top soil samples 1m apart to assess small scale reproducibility of the markers. The results of the study showed that the 16S, 18S and ITS markers appeared to be useful for forensic soil analysis since they were able to reliably discriminate contrasting soil samples.

The aim of this chapter is to examine the ability of 16S rRNA targeted sequencing to distinguish between similar and different soil types collected from

different locations. In order to achieve the outlined aim the following tasks were designed and executed:

- To collect soil samples from three different urban locations in Adelaide, including small locality replicates, and extract the DNA
- To analyse soil DNA specimens on the Ion Torrent platform to obtain high throughput sequencing data of the16S rRNA gene
- To undertake a bioinformatics treatment and analyses of the obtained HTS data followed by OTU profiles statistical comparison using hierarchical agglomerative clustering (CLUSTER) and non-metric multidimensional scaling (NMDS)
- To assess the discriminating power of the 16S rRNA targeted HTS approach towards distinguishing similar and different soil types taken from distinct urban locations.

## 3.2 Materials and methods

### 3.2.1 Soil sampling and DNA extraction

Four replicates of soil samples from sites A, B and C (Figure 3.1, for more details about soil sampling sites see Chapter 2, Materials and Methods Section) were collected 1 m apart as shown in Figure 3.1D, in winter season. In total 12 soil samples were collected for the subsequent analysis.

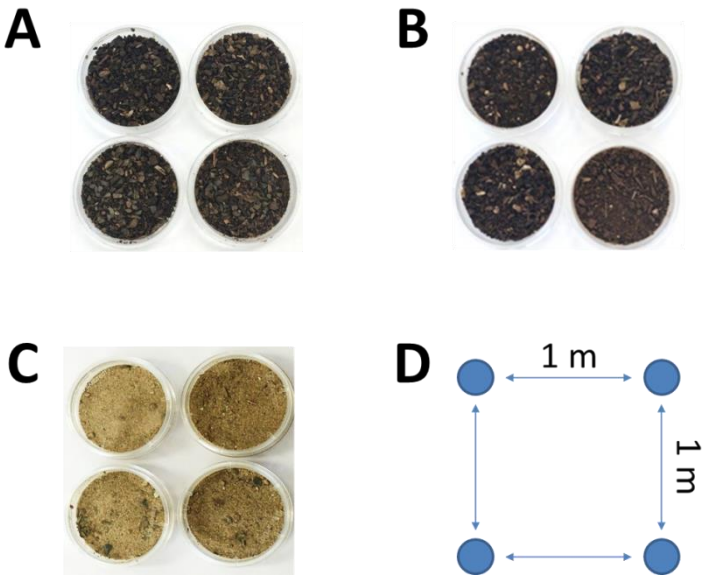



**Figure 3.1. A –** photograph of four replicative soil samples collected from location A (Flinders University), **B** – photograph of four soil samples collected from location B (Warradale reserve), **C** – photograph of four soil samples collected from location C (Brighton Esplanade), **D** – Schematic of replicate soil samples collection.

Soil samples were named according to their sampling sites, season (time) of sampling and number of replicates collected. Thus names of samples collected from location A start with a capital letter ‘A’, location B – ‘B’, location C – ‘C’; the next small letter represents season (time) of sample collection such as ‘a’ for autumn, ‘w’ for winter, ‘sp’ for spring and‘s’ for summer; a digit (from 1 to 4) represent a number of

replicate taken from the site. Thus, ‘Aw1’ means that soil sample was taken from location A at the winter time and is replicate number one.

DNA was extracted using the ZR soil DNA extraction kit from 0.05 g of soil.

### 3.2.2 PCR amplification and high throughput sequencing of the 16S rRNA gene

PCR amplification of the variable V3 region of 16S rRNA gene was performed using universal bacterial primers 341F and 518R (Muyzer 1993) tagged with Ion Torrent barcode and adapter sequences (Table 3.1).

**Table 3.1. Structure and sequence of the universal 16S rRNA gene specific primers used for PCR amplification.**

| Primer name | Primer sequence and structure |
|---|---|
| | ***Forward primers*** |
| | Primer A-key — Ion Torrent Barcode — *16S rRNA specific forward primer* |
| 341-F-10 | CCATCTCATCCCTGCGTGTCTCCGACTCAG—CTGACCGAAC—*cctacgggaggcagcag* |
| 341-F-11 | CCATCTCATCCCTGCGTGTCTCCGACTCAG—TCCTCGAATC—*cctacgggaggcagcag* |
| 341-F-12 | CCATCTCATCCCTGCGTGTCTCCGACTCAG—TAGGTGGTTC—*cctacgggaggcagcag* |
| 341-F-13 | CCATCTCATCCCTGCGTGTCTCCGACTCAG—TCTAACGGAC—*cctacgggaggcagcag* |
| 341-F-14 | CCATCTCATCCCTGCGTGTCTCCGACTCAG—TTGGAGTGTC—*cctacgggaggcagcag* |
| 341-F-15 | CCATCTCATCCCTGCGTGTCTCCGACTCAG—TCTAGAGGTC—*cctacgggaggcagcag* |
| 341-F-16 | CCATCTCATCCCTGCGTGTCTCCGACTCAG—TCTGGATGAC—*cctacgggaggcagcag* |
| 341-F-17 | CCATCTCATCCCTGCGTGTCTCCGACTCAG—TCTATTCGTC—*cctacgggaggcagcag* |
| 341-F-18 | CCATCTCATCCCTGCGTGTCTCCGACTCAG—AGGCAATTGC—*cctacgggaggcagcag* |
| 341-F-19 | CCATCTCATCCCTGCGTGTCTCCGACTCAG—TTAGTCGGAC—*cctacgggaggcagcag* |
| 341-F-20 | CCATCTCATCCCTGCGTGTCTCCGACTCAG—CAGATCCATC—*cctacgggaggcagcag* |
| 341-F-21 | CCATCTCATCCCTGCGTGTCTCCGACTCAG—TCGCAATTAC—*cctacgggaggcagcag* |
| 341-F-22 | CCATCTCATCCCTGCGTGTCTCCGACTCAG—TTCGAGACGC—*cctacgggaggcagcag* |
| 341-F-23 | CCATCTCATCCCTGCGTGTCTCCGACTCAG—TGCCACGAAC—*cctacgggaggcagcag* |
| | ***Reverse primer*** |
| | Primer P1-key — *16S rRNA specific reverse primer* |
| 518-R | CCTCTCTATGGGCAGTCGGTGAT—*attaccgcggctgctgg* |

PCR amplification was performed in 25 µL of total volume with 0.2 µM 341F and 0.2 µM 518R primers, 0.2 mM of dNTPs, 1 × HotStar buffer (Qiagen, Germany), 2.5 mM $MgCl_2$ and 0.625 U HotStar Taq DNA polymerase (Qiagen, Germany). Reaction conditions were 94 $^{o}$C for 15 min (initial denaturation) followed by 30 cycles of 1 min at 94 $^{o}$C (denaturation); 1 min at 53 $^{o}$C (primers annealing); 1 min at 72 $^{o}$C (primer extension) and a final elongation for 5 min at 72 $^{o}$C, as described previously (Jenkins et al. 2010). The obtained PCR products were purified using the QIAquick gel extraction kit (Qiagen, Germany) after preparative electrophoresis in 2% agarose gel stained by EtBr. Concentration of the PCR products was measured by a Qubit Fluorometer (Life Technologies) using the HS DNA quantification kit (Life technologies). Samples were pooled in equimolar concentrations for subsequent sequencing.

High throughput DNA sequencing was conducted using Ion Torrent platform (Life Technologies) at the Australian Genome Research Facility (AGRF, http://www.agrf.org.au/, Adelaide, SA, Australia).

The pooled purified PCR products were amplified by emulsion PCR onto Ion Sphere Particles (ISPs) using an Ion OneTouch 200 Template Version 2 DL Kit (Life Technologies) on an Ion One Touch machine (Life Technologies). The template ISPs were recovered from the emulsion, and the ratio of template ISPs to empty ISPs was determined by a fluorometric assay using fluorescently labelled oligonucleotides complementary to adapter sequences. The optimal template signal ratio was determined to be between 10% and 40%. Positive template ISPs were biotinylated during the emulsion PCR process, so that samples with an optimal template signal ratio were then enriched with Dynabeads MyOne Streptavidin C1 beads (Life Technologies) using an Ion ES robot (Life Technologies). Enriched ISPs were sequenced on an Ion 316 chip

using an Ion Torrent PGM sequencer (Life Technologies) and the Ion PGM 200 Sequencing Kit according to the manufacturer's instructions (Life Technologies). Torrent Suite software version 3.2 was used to parse the barcoded reads and to generate run metrics, including chip loading efficiency and total read counts and quality. Sequencing output represented separate FastQ files for each metagenomic DNA sample.

### 3.2.3 Quality filtering of the obtained sequencing data.

Cutadapt v1.1 software (Martin 2011) was used for primer trimming from the raw sequencing reads using strict zero mismatch threshold parameters both for 341F and 518R primers: -b [Forward; CCTACGGGAGGCAGCAG] –b [Reverse complement; CTGCTGCCTCCCGTAGG] and -b [Forward; ATTACCGCGGCTGCTGG] –b [Reverse complement; CCAGCAGCCGCGGTAAT], respectively. All reads with a length less than 100 bp were removed during the primer trimming process (parameter: –m [100 # discard reads that are shorter than min-length]).

Fastq_quality_filter tool (http://hannonlab.cshl.edu/fastx_toolkit) was used to remove the reads with a Phred quality score less than 20 for 90% of the read (parameters: -q 20 –p 90).

### 3.2.4 OTU picking and taxonomy assignment.

Quality filtered datasets were analysed using QIIME open source software version 1.7.0 (Caporaso et al. 2010). Sequences were clustered into OTUs using UCLUST (Edgar, 2010) open reference clustering protocol based on the default percent identity of 97% (Rosen et al. 2012). Then the number of sequences sharing 97% of identity in the created OTU clusters was counted. These counts became the abundance

data which were subjected to OTU-based statistical analysis. The OTU clusters having only one member, also called singletons, were discarded from the subsequent analysis as they were considered as chimeric sequences being produced due to sequencing artefacts (Haas et al. 2011). The May 2013 release of the Greengenes reference database (http://greengenes.secondgenome.com/downloads) was used for taxonomic annotation of the obtained OTUs. After taxonomic assignment QIIME generates an OTU abundance table as a BIOM file that can be used for a wide range of analyses (McDonald & Clemente 2012). The resulting OTU table was then rarefied at 79,067 sequences per sample (the minimal number of remaining reads in any of the samples (Table 3.2)).

### 3.2.5 Statistical analysis

Statistical approaches used in this chapter were the same as those described in Chapter 2. In brief, statistical analysis was performed in the Primer 6 package using CLUSTER and NMDS tools. ANOSIM analysis was used for the evaluation of significance of soil metagenomic OTU profiles differences.

### 3.2.6 Step-by-step procedure of the likelihood ratio (LR) model computation:

- A pair-wise Bray-Curtis similarity score matrix was constructed for all the samples' profiles included in the investigation and exported to MS Excel.
- The obtained similarity scores were divided into two groups: the first group included the scores from the comparisons of the profiles from the soil samples collected within a site; the second group included the scores from the comparisons of the profiles from the soil samples collected from different sites

- The average similarity score for each group (function AVERAGE) was calculated.
- The standard deviation based on the entire dataset, including within a site and between sites similarity scores was calculated. (function STDEV)
- The histograms of Bray-Curtis similarity scores for each group (Data/Analysis/Histogram in Excel) were built
- Assuming that similarity scores for each group have a normal distribution and assuming that these distributions for each group have equal variance, which is the variance calculated based on entire dataset, a model of probability distribution functions (PDFs) of similarity scores for within - and between-site groups (function NORM.DIST(Value,Mean,Standard_dev,FALSE)) was made.
- The LR values were calculated using these PDFs for every Bray-Curtis similarity score, which ranges from 0 to 100%. LR = NORM.DIST (Value,Mean(intra-group),Standard_dev,FALSE) / NORM.DIST(Value,Mean(inter-group),Standard_dev,FALSE)
- The discriminating power of the 16S rRNA sequencing approach and false positive and false negative rates were assessed using Tippett plot (Zadora et al. 2013).

## 3.3 Results and discussion

### 3.3.1 Soil sampling, DNA extraction and amplification.

Four replicative soil samples approximately 1 m apart from three locations A ($Aw_{1-4}$), B ($Bw_{1-4}$) and C ($Cw_{1-4}$) were collected in Adelaide in winter time.

After the extraction of DNA from soil samples (ZR soil DNA extraction kit), PCR amplification with primers specific to variable region V3 of 16S rRNA gene of Bacteria and Archaea was performed. The PCR products were then sequenced at the AGRF. The resulting sequencing datasets were named by the addition of a '16S' prefix in front of the soil sample name. For example, '16S-Bw3' means that the soil sample was taken from location B in winter, was replicate number three and was sequenced using the 16S rRNA targeted approach.

### 3.3.2 Primer trimming and quality filtering.

Sequencing of PCR amplified 16S rRNA gene fragments produced on average 363,325 (130,293 – 1,202,914) reads across all soil metagenomic samples. Primer trimming and quality filtering resulted in an average 44% of reads eliminated from each dataset.

**Table 3.2. General statistics of sequencing data processing. Yellow rows indicate datasets excluded from the subsequent analysis due to failed sequencing.**

| Sample | Initial N of reads | N of bp | After primers trimming | | After Quality Filtering | |
|---|---|---|---|---|---|---|
| | | | N of reads | % | N of reads | % |
| **16S_Aw1** | 166,671 | 22.5 | 114,029 | 68 | 79,067 | 47 |
| **16S_Aw2** | 145,003 | 21.9 | 118,049 | 81 | 84,052 | 58 |
| **16S_Aw3** | 156,012 | 24.9 | 138,116 | 89 | 105,964 | 68 |
| **16S_Aw4** | 130,293 | 20.6 | 112,695 | 86 | 80,590 | 62 |
| **16S_Bw1** | 294,086 | 43.9 | 237,116 | 81 | 165,079 | 56 |
| **16S_Bw2** | 104 | | 55 | 53 | 37 | 36 |
| **16S_Bw3** | 208,599 | 31.7 | 172,171 | 83 | 124,868 | 60 |
| **16S_Bw4** | 249,821 | 37.5 | 201,947 | 81 | 143,863 | 58 |
| **16S_Cw1** | 1,202,914 | 157.3 | 784,758 | 65 | 534,881 | 44 |
| **16S_Cw2** | 508,518 | 84.9 | 477,468 | 94 | 262,840 | 52 |
| **16S_Cw3** | 1,083 | | 782 | 72 | 484 | 45 |
| **16S_Cw4** | 571,329 | 83.3 | 444,351 | 78 | 319,065 | 56 |
| **Average** | 363,325 | 52.9 | 280,070 | 81 | 190,027 | 56 |
| **Min** | 130,293 | 20.6 | 112,695 | 65 | 79,067 | 44 |
| **Max** | 1,202,914 | 157.3 | 784,758 | 94 | 534,881 | 68 |

### 3.3.3 Taxonomic analysis

The QIIME toolkit (Caporaso et al. 2010) was used to analyse the results by efficient OTU selection and taxonomy assignment. Of the classified reads *Actinobacteria* and *Proteobacteria* were the dominant Phyla across all soils, which together accounted for approximately 80% of the reads classified for A and B soil samples and 65% for soil sample C. *Acidobacteria*, *Bacteroidetes*, *Chloroflexi*, and *Verrucomicrobia* taxa each comprised less than 5% of classified reads. Of note *Cyanobacteri*a (12%), *Firmicutes* (4%*)* and *Gemmatimonadetes* (2%) were prevalent in the sequence datasets from the location C soil samples (Figure 3.2), but represented less than 1% of classified reads in the datasets from soils A and B.

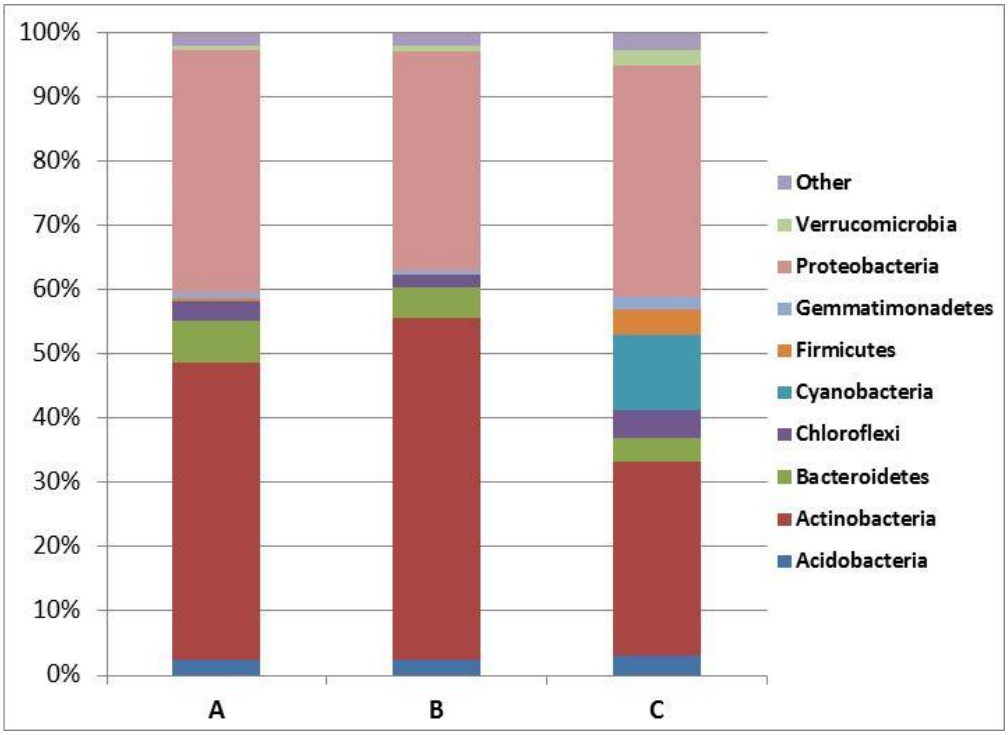



**Figure 3.2. Average relative abundances of taxa at the phylum level, identified by 16S rRNA targeted sequencing of DNA extracted from soil samples taken from three geographically distinct locations (A, B and C).**

### 3.3.4 Comparison of soils based on their OTU profiles

Comparison of the 16S rRNA OTU profiles was conducted by unconstrained multivariate statistical analyses, namely NMDS ordination analysis was used in conjunction with CLUSTER analysis (Figure 3.3). Pairwise Bray-Curtis similarity scores were calculated for all samples using square-root transformed data of the OTUs relative abundances. CLUSTER analysis successfully grouped all samples according to their collection sites with average profile similarity of 61.0 ± 2.9%, 67.3 ± 0.1% and 52.6 ± 6.0% for soils from sites A, B and C, respectively (Figure 3.3A, Table 3.3).

**Table 3.3. Average Bray-Curtis similarity scores obtained from the pairwise comparison of 16S OTU-based soil profiles.**

| Sample name | 16S_A | 16S_B | 16S_C |
|---|---|---|---|
| **16S_A** | 61.0 ± 2.9 | | |
| **16S_B** | 53.7 ± 3.3 | 67.3 ± 0.1 | |
| **16S_C** | 36.0 ± 4.9 | 32.9 ± 5.0 | 52.6 ± 6.0 |
| **Within a site** | 60.5 ± 6.3 | | |
| **Between sites** | 41.6 ± 10.3 | | |
| **SD (full data)** | 12.6 | | |

Data represented as average ±SD between profiles from replicative samples taken from each site

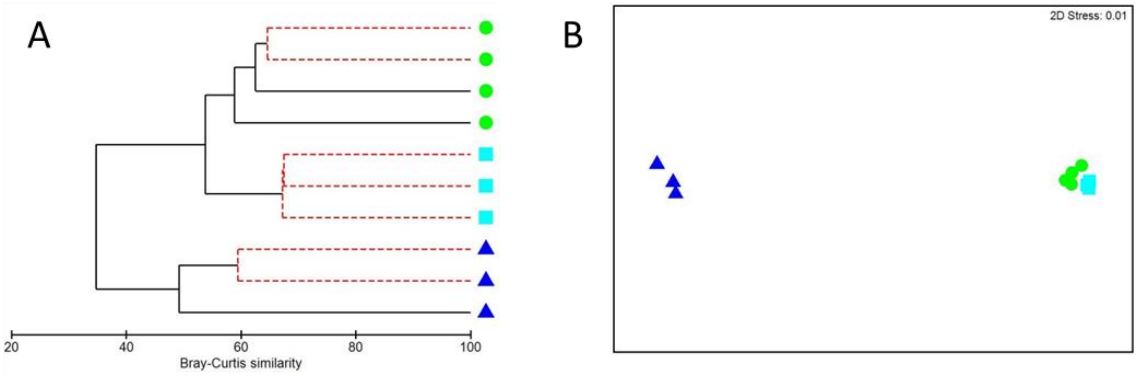



**Figure 3.3. Comparison of the 16S OUT-based soil profiles from three geographically distinct locations A (●), B (■) and C (▲).** A Bray-Curtis similarity matrix was calculated from the square-root transformed data of the 16S OUT abundance profiles. The Bray-Curtis matrix was used for generating a CLUSTER dendrogram and NMDS ordination plot. CLUSTER analysis (A) Red dashed branches on the CLUSTER dendrogram indicated no significant difference between metagenomic profiles (supported by the SIMPROF analysis, $p<0.05$). NMDS unconstrained ordination (B). The NMDS plot displayed distances between samples. Data points that are closer to each other represent samples with highly similar metagenomic profiles.

SIMPROF analysis confirmed the formation of a genuine cluster only for the samples from location B (Figure 3.3A, square symbols). NMDS ordination in turn displayed two separate groupings, where the first group consisted of the samples from location C only, and the second one was divided into two sub-clusters made of samples from sites A and B. A low stress level of 0.01 showed no loss of information occurred during projection of the analysis output in the multidimensional scale on the two-dimensional NMDS ordination plot. Importantly these distances between the samples reflect the nature of the soil types investigated indicating that visually similar soils from

locations A and B share more similarity compared with any of two with soil C. It is worth noting that differences between the groups tested by the ANOSIM analysis were shown to be statistically significant with Global R = 0.877 and $p$ = 0.0005. This result confirms previous findings (Young et al. 2014) that HTS-based 16S rRNA gene survey allows for reliable discrimination of contrasting/different soil types.

### 3.3.5 A LR-model for soil discrimination

To discuss the results it is appropriate to apply a LR Baysian statistics. The Bayesian framework (using LR framework) is commonly used for the interpretation of evidence in forensic casework (Aitken & Taroni 2004). It has been applied, amongst other things, to DNA interpretation (Christopher M. Triggs, John S. Buckleton 2004), forensic voice comparison (Gonzalez-Rodriguez 2006), forensic face recognition (Ali et al. 2011), and physicochemical evidence interpretation (Zadora et al. 2013).

The Bayes theorem (equation 1) shows how new evidence (the LR) can change the prior odds (prior background knowledge) resulting in the posterior odds used by jurors to make their decision after observing the evidence.

$$\underbrace{\frac{P(H_o)}{P(H_1)}}_{\text{Prior odds}} \times \underbrace{\frac{P(x|H_0)}{P(x|H_1)}}_{\substack{\text{Likelihood} \\ \text{Ratio (LR)}}} = \underbrace{\frac{P(H_0|x)}{P(H_1|x)}}_{\substack{\text{Posterior} \\ \text{odds}}} \quad (1)$$

Non-scientific information constitutes the prior odds, while the forensic scientist is responsible for the quantitative evaluation of the evidence ($x$) in the form of a likelihood ratio (LR, equation 2). The LR is defined as a ratio of the probabilities of the evidence ($x$) given each of two competing hypotheses: $H_0$ – the hypothesis that the two

samples have the same origin, versus $H_1$ – the hypothesis that they have different origins.

$$LR(x) = \frac{P(x|H_0)}{P(x|H_1)} \quad (2)$$

The LR ratio has a range from 0 to infinity. All LR values >1 support the $H_0$ hypothesis, while LRs < 1 support the $H_1$ hypothesis. The LR value can be translated into a verbal equivalent when presented in court to support the strength of the evidence (Table 3.4). This chapter describes a method of applying an LR estimation to a comparison of soil metagenomic profiles.

**Table 3.4. Verbal equivalent of likelihood ratio (LR).**

| Verbal equivalent* | $Log_{10}$(Likelihood Ratio) |
|---|---|
| Very strong support $H_0$ | > 14 |
| Strong support $H_0$ | 3 to 4 |
| Moderately strong support $H_0$ | 2 to 3 |
| Moderate support $H_0$ | 1 to 20 |
| Limited support $H_0$ | 0 to 1 |
| Limited support $H_1$ | 0 to -1 |
| Moderate support $H_1$ | -1 to -2 |
| Moderately strong support $H_1$ | -2 to -3 |
| Strong support $H_1$ | -3 to -4 |
| Very strong support $H_1$ | < -4 |

* A verbal interpretation is based on the ideas of Evett et.al.(Evett et al. 2000) .

### 3.3.6 Scoring method of LR computation.

Figure 3.4 shows the main steps towards the calculation of the likelihood ratio (LR) for evidence interpretation based on 'simulated' experimental data of soil metagenomic profile comparisons. The Bray-Curtis similarity scores obtained by pair-wise comparison of the soil metagenomic profiles are considered as evidence (*x*). In this study a set of Bray-Curtis similarity scores, obtained by comparing soil metagenomic profiles from the same source, represents within site variability (WS group). Between

sites variability (BS group) is based on the Bray-Curtis similarity scores derived from the comparison of soil metagenomic profiles from different locations.

This study shows that the soil metagenomic profiles derived from the soil samples collected within a site had more similarity than those collected from different sites. The average similarities of the profiles from samples collected within a site and from different sites were 60.5 ± 6.3% and 41.6 ± 10.3%, respectively (Table 3.3).

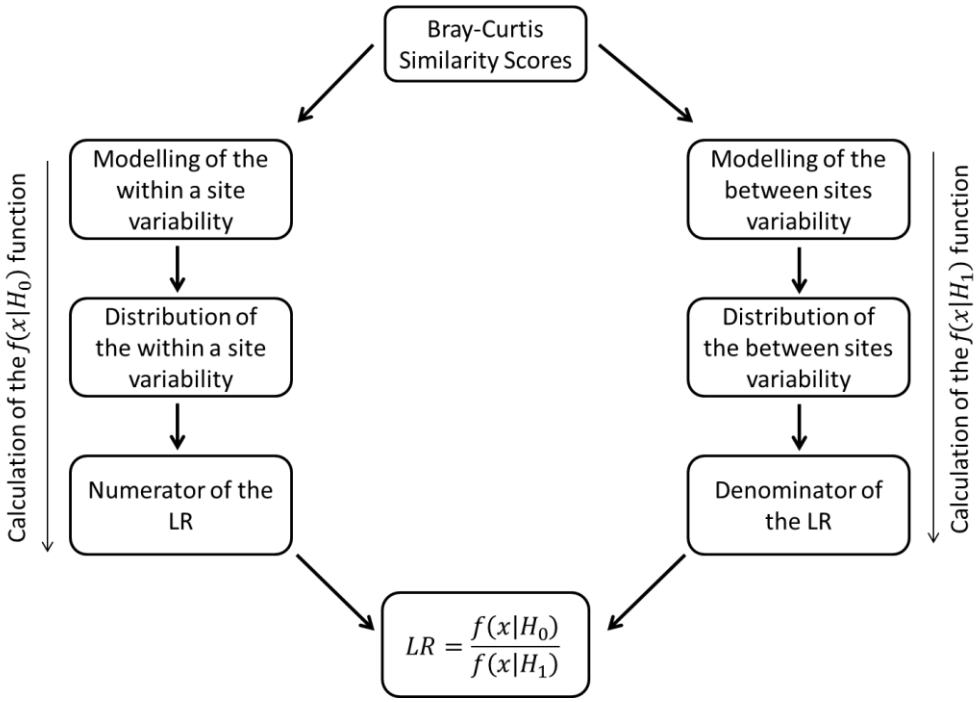



**Figure 3.4. Main steps of LR computation.**

The first step is to model WS and BS score distributions (Figure 3.5).

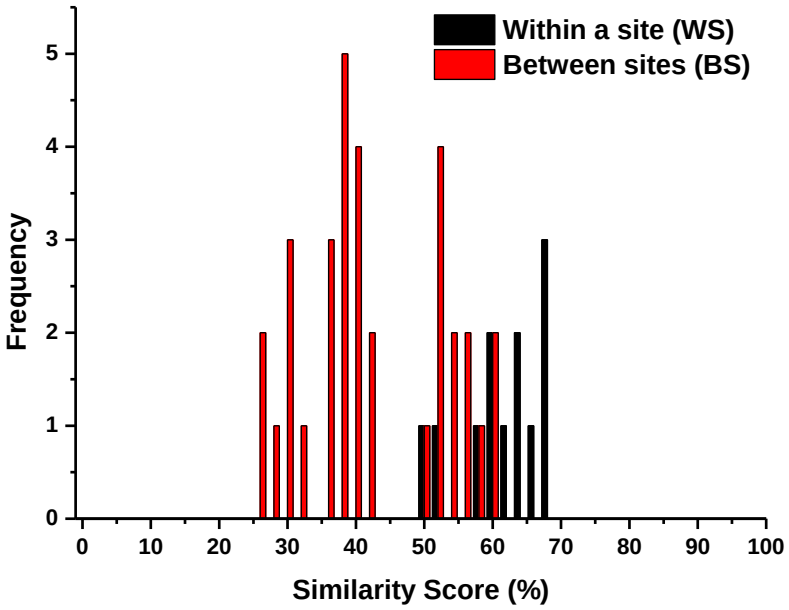



**Figure 3.5. Bray-Curtis similarity scores distribution for within site (black bars) and between site groups (red bars).**

In the current study the number of samples is low so two simplifying assumptions were made in order to model the similarity score variability. The first assumption is that the similarity score distribution for each group (WS and BS) is described by a Gaussian model. The second assumption is that the two groups have the same variance (SD) as shown in Table 3.3. It is likely that these assumptions will be found to be not true with more intensive sampling. The two distributions are represented as probability density functions (pdf) of the scores under hypotheses $H_0$ (the pair of soils that produced score *x* originate from the same site) and $H_1$ (the pair of soils that produsced score *x* originate from different sites (Figure 3.6).

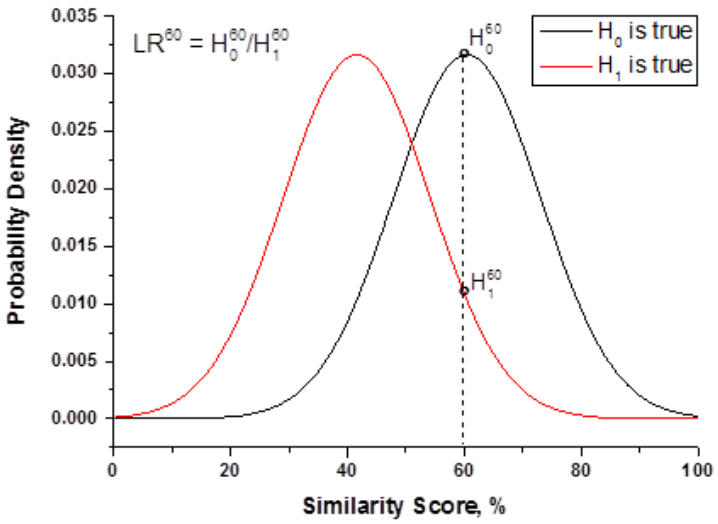



**Figure 3.6. Gaussian distribution of Bray-Curtis similarity scores for the within site group (black lines) and between sites group (red lines).**

Equation 3 shows the calculation of a likelihood ratio (LR) by using the probability density functions of the within site (WS) and between site (BS) distributions to give a numerical value for the evidence ($x$):

$$LR(x) = \frac{f(x|H_0)}{f(x|H_1)} \quad (3),$$

where $f(x|H_0)$ and $f(x|H_1)$ represent the WS probability density function and BS probability density function, and $x$ is the Bray-Curtis similarity score obtained by comparison of the soil metagenomic profiles.

Similarity scores range from 0 to 100%. For every similarity score within the range a likelihood ratio value was then calculated (Equation 3) and plotted as a function of the corresponding score (Figure 3.7).

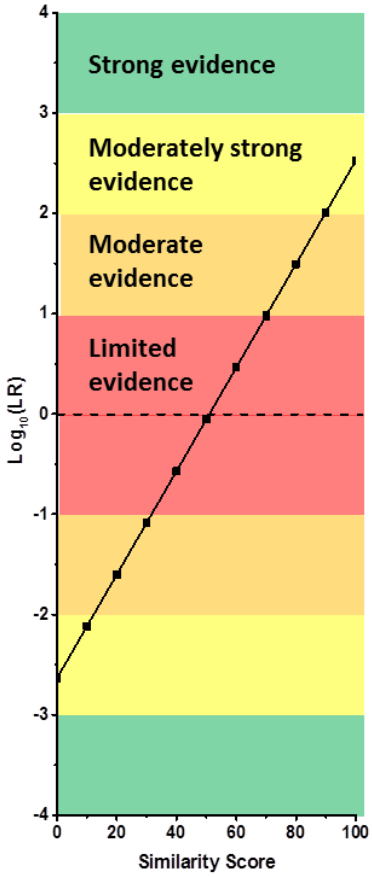



**Figure 3.7. Correlation of the decimal logarithm of LR ratios versus Similarity Scores obtained by comparing soil metagenomic profiles produced by 16S rRNA sequencing.**

It can be seen that the LR line intercepts $y = 0$ ($Log_{10}(LR) = 0$) at $x = 50\%$. This means that the probability of obtaining similarity score above 50% is more likely if two soil samples derived from the same origin ($H_0$ is true) than if they were collected from different sites ($H_1$ is true). As the LR value increases so does the support for $H_0$ over $H_1$. For example, with a similarity score of 90% between two soil samples it is 100 ($Log_{10}(LR) = 2$) times more likely that these soils have come from the same site. Alternatively, if the similarity score observed falls below 50% then it is more likely that these soils have derived from different sites. Such a representation of the association between similarity scores and the likelihood ratio values can be used as a 'calibration curve'. This will aid in transformation of new Bray-Curtis similarity scores obtained

from the comparison of new datasets into the LR values. However, before that the proposed model should be rigorously validated on a significantly larger number of varied soil samples from different locations treated the same way.

### 3.3.7 Discriminating power of 16S rRNA metagenomic sequencing for the analysis of similar and different soil types from different locations.

The next step was to estimate the discriminating power of the 16S rRNA HTS sequencing method being tested with regard to differentiating of soil samples taken from geographically different locations. In order to do this an estimation of the $Log_{10}$(LR) values for every pairwise comparison of soil samples included in the study was performed based on the proposed LR model (Figure 3.7). The $Log_{10}$(LR) values corresponding to scores obtained from the comparison of visually similar soils and different soils sets are presented in Table 3.5.

For the same sets of samples, i.e. visually similar and different soils from distant sites, probability density functions (pdf) of the obtained $Log_{10}$(LR) values were represented on probability distribution plots with respect to each hypotesis $H_0$ and $H_1$ for visually similar soils (Figure 3.8A); $H_0$' and $H_1$' for contrasting soils (Figure 3.8B,). For the ideal case scenario, a complete separation of distributions is expected if all scores from within site comparisons are characterised by $Log_{10}(LR) > 0$ which in turn supports the $H_0$ hypothesis (or $H_0$'). At the same time all scores from between sites comparisons should have only negative $Log_{10}$(LR) values, supporting the $H_1$ hypothesis (or $H_1$'). As can be seen on both plots (Figure 3.8A and Figure 3.8B) there is a degree of overlap between $Log_{10}$(LR) probability distributions supporting the competing propositions which indicates that some soils are being falsely identified as matching

while other soils are being falsely excluded as matching. Given these two distributions it is possible to estimate the siginificance of the LR model generated based on the experimental data obtained in the current study. A percentage of LR values for the current experimental data supporting the wrong proposition indicates false positive and false negative error rates which are a reflection of the value of the model and its discriminating power.

**Table 3.5. $Log_{10}$(LR) values derived from Bray-Curtis similarity scores.**

| **Comparison of visually similar soils A & B** | | | | **Comparison of visually different soils A & B & C** | | | |
|---|---|---|---|---|---|---|---|
| Within a site A-A & B-B | | Between sites, A-B | | Within a site A-A & B-B & C-C | | Between sites A-C & B-C | |
| Similarity Score, % | $Log_{10}$(LR) | Similarity Score, % | $Log_{10}$(LR) | Similarity Score, % | $Log_{10}$(LR) | Similarity Score, % | $Log_{10}$(LR) |
| 59 | 0.41 | 52 | 0.05 | 59 | 0.41 | 37 | -0.72 |
| 61 | 0.51 | 54 | 0.15 | 61 | 0.51 | 37 | -0.72 |
| 57 | 0.31 | 50 | -0.05 | 57 | 0.31 | 28 | -1.19 |
| 63 | 0.62 | 58 | 0.36 | 63 | 0.62 | 39 | -0.62 |
| 65 | 0.72 | 59 | 0.41 | 65 | 0.72 | 39 | -0.62 |
| 67 | 0.82 | 57 | 0.31 | 59 | 0.41 | 29 | -1.14 |
| 67 | 0.82 | 55 | 0.20 | 48 | -0.16 | 38 | -0.67 |
| 67 | 0.82 | 52 | 0.05 | 67 | 0.82 | 40 | -0.57 |
| | | 54 | 0.15 | 67 | 0.82 | 39 | -0.62 |
| | | 54 | 0.15 | 62 | 0.57 | 35 | -0.83 |
| | | 52 | 0.05 | 50 | -0.05 | 37 | -0.72 |
| | | 48 | -0.16 | 67 | 0.82 | 37 | -0.72 |
| | | | | | | 28 | -1.19 |
| | | | | | | 40 | -0.57 |
| | | | | | | 30 | -1.09 |
| | | | | | | 41 | -0.52 |
| | | | | | | 31 | -1.03 |
| | | | | | | 34 | -0.88 |
| | | | | | | 34 | -0.88 |
| | | | | | | 26 | -1.29 |
| | | | | | | 26 | -1.29 |

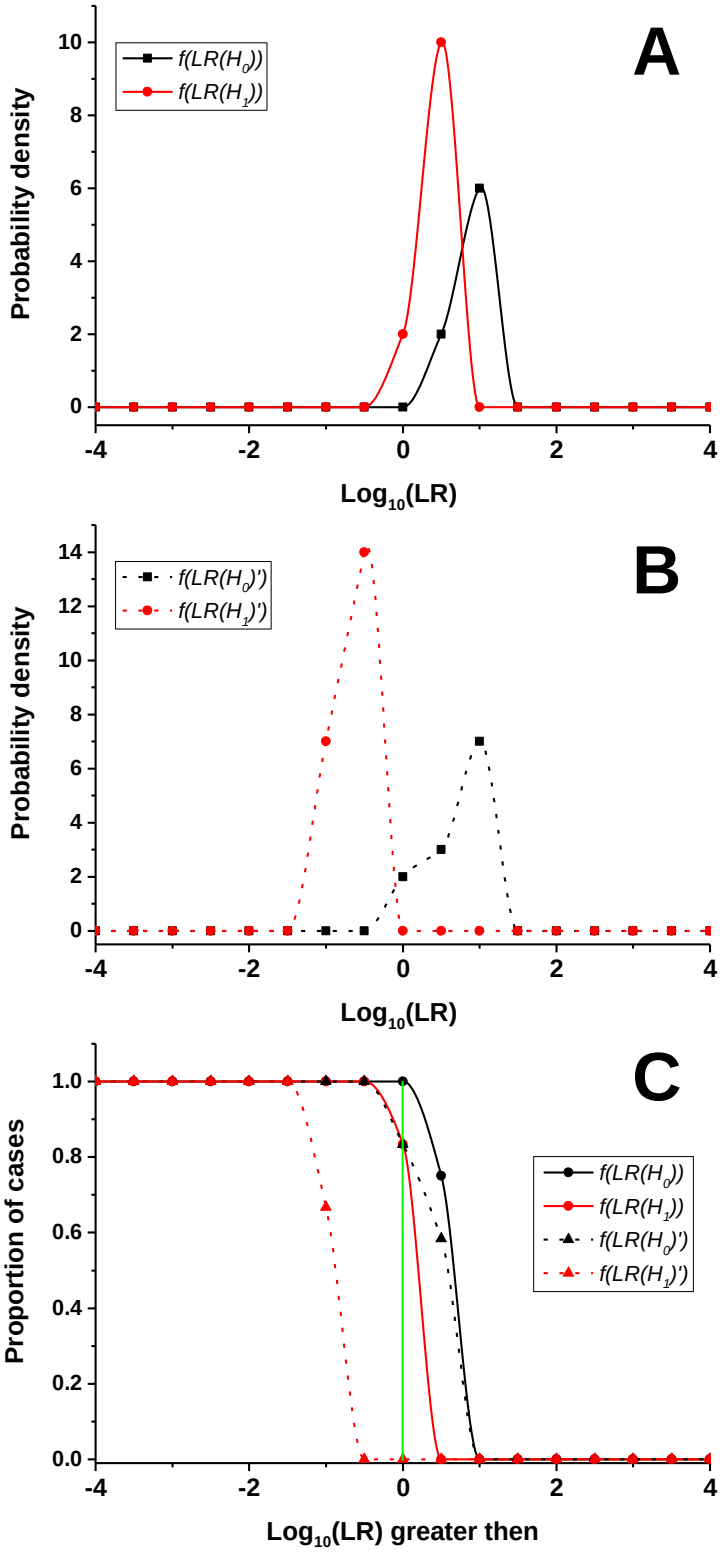



**Figure 3.8. Estimated probability density functions (pdfs) of $Log_{10}$(LR) values for (A) visually similar soils, (B) different soils. (C) Tippett Plot.**

The cumulative version of these $Log_{10}$(LR) probability distributions which is used to represent how many cases from the current experimental data are above a given value of LR with respect to each proposition is called a Tippett plot (Zadora et al. 2013). The Tippett plot (Figure 3.8C) aids visualisation of the false positive/negative rates of the proposed model. Each of the curves represents the inverse cumulative proportion of the $Log_{10}$(LR) values supporting the competing propositions $H_0$ and $H_1$ ($H_0$' and $H_1$') shown in the Figure 3.8C as black and red traces (dashed black and dashed red traces respectively). The rates of false positive and false negative errors are visible in the Tippett plot at the intersection of each of the curves for either LR within site ($H_0$ and $H_0$') or LR between sites ($H_1$ and $H_1$') and the imaginary line (Figure 3.8C, green solid line) going vertically through value zero on the *X* axis ($Log_{10}$(LR) = 0). Thus in the comparison of visually similar soils the false negative rate is 0% (Figure 3.8C, black solid line), and the false positive rate is 0.83 (83%). In the case of the comparison of entirely different soils, the false negative rate is 0.17 (or 17%), and the false positive rate is 0%. Given the obtained false positive and false negative errors rates it can be concluded that 16S rRNA sequencing approach allows for discrimination of the soil metagenomes from contrasting habitats (different land management and vegetation types) however it does not reliably discriminate similar looking soils taken from different sites as indicated by the high rate (83%) of false positive error.

## 3.4 Conclusion

DNA typing of the microbial community based on high throughput DNA sequencing of the 16S rRNA bacterial marker is a gold standard method of microbial community characterisation in environmental ecological research, highlighted by a variety of publications in peer-reviewed scientific journals (Tringe & Hugenholtz 2008; Caporaso & Lauber 2011; Mande et al. 2012b; Valverde & Mellado 2013).

In the current study the ability of 16S gene sequencing is assessed with regard to discriminating soils that are both visually similar and entirely different, sourced from three urban locations in Adelaide, South Australia. Sequencing of the V3 region of the 16S rRNA gene was performed according to standard procedure followed by OTU-based bioinformatic analysis of the obtained metagenomic data using default settings within QIIME open source software. Multivariate statistical analysis such as CLUSTER and NMDS ordination was then performed for the comparison of the OTU profiles generated. CLUSTER analysis revealed that all samples, including replicates from each sampling site, grouped correctly according to the sampling sites. NMDS ordination analysis also demonstrated correct clustering of samples, indicating that bacterial 16S profiles from visually similar soils (A and B) have more similarity than different soils.

In order to ‘translate’ the results of the soil metagenomic profiles comparisons to the terminology accepted by the forensic community and the courts, a Likelihood ratio model was applied. Modelling the data by the proposed manner allows the assessment of the significance of similarity scores determined from soil samples comparisons. Specifically, the ratio of the WS and BS probability density functions provides a measure of the probability of obtaining the observed similarity score if the soil samples come from the same site as opposed to different sites. This study is a pilot and relies on

a limited number of samples available for analysis. Increasing the number of samples may result in corrections to the distribution of the Bray-Curtis similarity scores and would alter the trends seen here. Further, the proposed model for transformation of soil metagenomic profile similarity scores into LR values supporting or opposing the proposition that two soil samples have originated from the same site (Figure 3.7) can be used for validation based on the new unrelated data gathered.

The use of the Bayesian framework allows the Likelihood ratio presented by the scientist to be assessed with the prior odds and gives the court some assistance in determining the posterior odds that two soils have a common origin. It clearly places the scientist in the domain of considering the evidence given a hypothesis which can readily translate to different scenarios or across difference evidence types. Its application to soil DNA typing would be to compare the competing hypotheses; is the unknown soil sample a match to samples taken from the scene of a crime or does the soil sample originate from another unrelated site. The advantage of this framework is the ability to change the hypothesis according to different prosecution or defence hypotheses given.

The other very important outcome of this study is that the discriminating power of the 16S DNA sequencing approach was estimated and false positive and false negative rates were assessed for discrimination of visually similar and different soils. The CLUSTER and NMDS statistical tools showed correct clustering of similar looking soils samples A and B. However, by using a Likelihood ratio approach a high level of false positive errors were identified (83%). The analysis of visually different soil types gave a lower rate of false positives and false negatives, 0% and 17% respectively.

It is envisaged that both the increase in the number of analysed soil samples and spreading their geographical origin, along with testing of other genetic markers within

the ribosomal operon will help not only better assess the performance of targeted metagenomics at discrimination of visually similar and different soils collected from distinct locations, but will also fine-strenghten the model that has been developed.

# Chapter 4. Arbitrary primed PCR based sequencing of soil metagenome for forensic soil discrimination

## 4.1 Introduction

Advances in high throughput DNA sequencing technologies represent a leap forward for forensic comparisons of soils based on the analysis of microbial community composition. Preliminary results on forensic analysis of soils using HTS sequencing of phylogenetic markers have been reported recently (Lilje et al. 2013; Giampaoli et al. 2014; Young et al. 2014). Previously, Waters *et al*. had proposed an original approach which relies on utilizing arbitrarily primed PCR amplification (AP-PCR) (Welsh & McClelland 1990) for forensic soil DNA typing (James M. Waters, Graham Eariss, P. Jane Yeadon, K. Paul Kirkbride 2012). The authors reported that the results of AP-PCR were successfully evaluated using hybridisation on custom-made microarrays. This was the first reported example of the application of high throughput DNA screening technology based on DNA microarrays for forensic soil discrimination. However, the authors also suggested using AP-PCR in conjunction with high throughput DNA sequencing for the most powerful soil DNA comparative analysis.

AP-PCR belongs to multiple arbitrary amplicon profiling (MAAP) methods, which use PCR amplification of DNA with a single primer of arbitrarily chosen sequence (Caetano-Anolles 1994). These methods also include random amplified polymorphic DNA (RAPD) (Williams et al. 1990) and DNA amplification fingerprinting (DAF) (Caetano-Anolles 1993) (Table 4.1). These fragments are then separated by electrophoresis in acrylamide or agarose gels producing a DNA fingerprint. Discriminating between and linking DNA fingerprints from soils and microbial communities are based on the scoring of ‘present’ or ‘absent’ bands (Franklin et al. 1999; Srinivasiah et al. 2013; Vettori et al. 1996; Wikström et al. 1999).

Additionally, individual or total PCR fragments from the DNA fingerprints could be cloned and sequenced to be used further in genetic composition studies. For example, such RAPD based amplicon sequence analysis was employed for the evaluation of soil microbial communities and its biological functions and metabolic pathways (Amorim et al. 2012). Structure and seasonal dynamic changes of viral assemblages persisting in aquatic sediments were also examined by RAPD amplification followed by cloning and Sanger sequencing (Helton & Wommack 2009).

**Table 4.1. General characteristics of multiple arbitrary amplicon profiling (MAAP) techniques.**

| Characteristics | RAPD | AP-PCR | DAF |
|---|---|---|---|
| Primer length (nt) | 9-10 | 18-32 | 5-15 |
| Annealing temperature ($^{o}$C) | 35-42 | 35-50 | 10-65 |
| Amplification stringency | low | low to high | low to high |

The rationale for using the MAAP amplification techniques is that no prior knowledge of the target sequence is required, minute amounts of DNA can be amplified and during the amplification process a massive number of random loci of all the metagenome constituents is examined regardless of their taxonomic origin. As the main feature and limitation of soil forensic evidence samples is a small initial size of the sample, it is therefore likely that the DNA material obtained from these samples will need to be amplified to allow analysis. The ability to analyse the entire genetic composition is desirable for forensic science as it provides more information for comparison and differentiation between soil samples and can result in the identification of the unique/signature features of soil. MAAP techniques represent a good choice as a DNA-based method for forensic soil discrimination.

It should be noted that the lack of reproducibility of the RAPD method (and related assays) in some studies has led investigators to report that the RAPD fingerprinting method was not reliable (Khandka et al. 1997). Several factors have been shown to influence RAPD profiles, including concentration of primer, deoxynucleotide triphosphates (dNTPs), DNA polymerase and $MgCl_2$, DNA template quality and thermal cycling conditions (Tyler et al. 1997). Obviously all these parameters play an important role, and one should clearly understand before beginning any experiments the impacts and effects of such parameters on RAPD profiling. Rigorous optimization studies of RAPD (and related AP-PCR reaction) were undertaken in order to improve the consistency, reproducibility and reliability of these assays (Ashayeri-Panah et al. 2012; Atienzar & Jha 2006; Dabrowski et al. 2003; Jhang & Shasany 2012; Vickery et al. 1998).

MAAP techniques are not new to forensic science. Some of the earliest studies using RAPD allowed for reliable discriminating of different seeds, seedlings, leaves and flowerheads (marijuana) of *Cannabis sativa* (Jagadish et al. 1996; Gillan et al. 1995). Application of the RAPD analysis was also demonstrated for differentiation of Papaver species (Shoyama et al. 1998).

The aim of the proof-of-concept study presented in this chapter is to investigate the potential of single arbitrary primed amplification (AP-PCR) coupled with subsequent HTS DNA sequencing (AP-PCR based sequencing) as a tool to examine the entire soil genetic composition for forensic soil discrimination. To achieve this goal, the following tasks were designed and performed:

- Three different urban sites in the Adelaide area were sampled and the total DNA was extracted;

- AP-PCR amplification was optimised with particular regard to primer annealing temperature, $Mg^{2+}$ concentration and amount of initial DNA template;
- instrumental reproducibility of the AP-PCR based sequencing technique for the generation of repeatable metagenomic profiles from the same soil samples was assessed;
- a small scale study of spatial variability and seasonal changes of soils genetic compositions were evaluated;
- An evaluation of the discriminating power of the AP-PCR based sequencing technique for visually similar soils and different soil types was undertaken using a Bayesian Likelihood ratio approach.

## 4.2 Materials and methods

### 4.2.1 Soil sampling

Soil samples were collected at three urban sites within the Adelaide metropolitan area approximately 3 km apart (for detailed characteristics of soil sampling sites see Chapter 2, Materials and Methods section). Four replicates for each soil sample from sites A, B and C were collected 1 m apart in winter season (for details see Chapter 3, Materials and Methods section). These samples constituted Set 1 in the current chapter. For the study presented in the current chapter additionally one soil sample from each of the sampling sites was taken at three additional time points corresponding to different seasons during the year. These soil samples were allocated in Sets 2, 3, and 4 (Table. 4.2). In total 21 soil samples were collected and analysed in the scope of the current study.

**Table 4.2. Notation of soil replicate samples collected from A, B and C locations.**

| | GPS coordinates | Set 1 winter | Set 2 spring | Set 3 summer | Set 4 autumn |
|---|---|---|---|---|---|
| **Flinders University (A)** | S35 01 43.42 E138 34 16.26 | Aw1-4 | Asp | As | Aa |
| **Warradale reserve (B)** | S35 00 58.09 E138 32 12.03 | Bw1-4 | Bsp | Bs | Ba |
| **Brighton Esplanade (C)** | S35 02 13.17 E138 51 59.22 | Cw1-4 | Csp | Cs | Ca |

Samples are named according to their sampling location, season (time) of sampling and number of replicates. Thus names of samples collected from location A starts with a capital letter 'A', location B – 'B', location C – 'C'; the next small letter represents season (time) of sample collection such as 'a' for autumn, 'w' for winter, 'sp' for spring and 's' for summer; a following digit (from 1 to 4), if any, represent a number of replicates taken from the location, absense of the digit shows that only one sample was collected from the sampling site.

### 4.2.2 DNA extraction

Metagenomic DNA was isolated from 0.05 g of each soil sample within 24 h after the collection using ZR Soil Microbe DNA Kit (Zymo Research, USA) following the manufacturer's recommendations. The quality of the DNA extracts was verified by gel electrophoresis in a 1% agarose gel stained with ethidium bromide. DNA concentrations were determined using a Qubit dsDNA HS Assay Kit (Invitrogen, USA) on a Qubit fluorometer (Life technologies, USA). The obtained DNA extracts were stored at -20 $^{o}$C until analysis.

### 4.2.3 Arbitrarily primed PCR amplification

Arbitrarily designed oligonucleotide primers designated Seq5 and Seq5-RC (RC-reverse complement) (Waters et al. 2012), previously reported as being effective for generating AP-PCR soil metagenomic fingerprints, were used in this study. Additionally two pairs of complementary 18 nt long primers were randomly generated with a GC content of 60% (Table 4.3). All the primers were synthesised by Integrated DNA Technologies (USA).

**Table 4.3. Sequences, melting points and GC content of primers used for AP-PCR amplification.**

| Primer Name | Sequence 5' – 3' | Tm, $^{o}$C | GC, % |
|---|---|---|---|
| **Seq5*** | CCC TCG AAC ACC ACC TCC | 57.9 | 66.7 |
| **Seq5-RC (P5)*** | GGA GGT GGT GTT CGA GGG | 57.9 | 66.7 |
| **Seq6 (P6)** | GAG ATT GAC CTG CAC GCC | 56.4 | 61.1 |
| **Seq6-RC** | GGC GTG CAG GTC AAT CTC | 56.4 | 61.1 |
| **Seq7** | AAT CAC CCC TGC TCC CGT | 59.2 | 61.1 |
| **Seq7-RC** | ACG GGA GCA GGG GTG ATT | 59.2 | 61.1 |

* Primers were derived from previous investigations of Waters *et al*. (Waters et al. 2012).

Amplification of extracted soil DNA was performed with primers Seq5-RC (P5) and Seq6 (P6) (Table 4.2). As a template, 4 ng of metagenomic DNA extracts was used. The 25 μL final reaction volume contained 1×HotStar Taq buffer (Qiagen, Germany), 2.5 mM $Mg^{2+}$, 0.2 mM of each dNTPs, 0.4 μM of the arbitrary chosen primer, and 0.625 U HotStarTaq DNA polymerase (Qiagen, Germany). An initial 15 min denaturation step at 95 $^{o}$C was followed by 42 cycles of 30 s at 94 $^{o}$C, 30 s at 55 $^{o}$C and 1 min at 72 $^{o}$C. A final extension step of 7 min at 72 $^{o}$C was used to complete the reaction. The quality of amplification products was determined by 1% agarose gel electrophoresis and by quantification on a Qubit fluorometer (Life technologies, USA) after purification with a QIAquick PCR Kit (Qiagen, Germany).

### 4.2.4 Library preparation and sequencing

Library preparation and sequencing were performed for each sample at the Australian Genome Research Facility (AGRF, http://www.agrf.org.au/, Adelaide, SA, Australia) and Australian Cancer Research Foundation (ACRF) Cancer Genomics Facility (http://centreforcancerbiology.org.au/acrf/, Adelaide, SA, Australia) using Ion Torrent technology (Ion Torrent PGM Sequencer; Life Technologies, USA).

For the library preparation, 100 ng of AP-PCR amplification product from each soil sample was used. Amplification products were sheared using the Ion Shear Version 2 Kit (Life Technologies), aiming for a mean fragment size of ~200 bp. Each sample of sheared DNA was cleaned using AMPure beads (Agencourt) and a sub-sample of each was checked for size and concentration on an Agilent Tapestation, using a High Sensitivity DNA Tapescreen (Agilent). Each sample was then ligated to adapters from the Ion Xpress Barcoded Adapters 1–16 Kit using the Ion Plus Fragment Library Kit

according to the manufacturer's instructions (Life Technologies). After ligation, products underwent nick-translation, clean up using AMPure beads and additional library amplification by PCR. Following library amplification, two rounds of clean up using AMPure beads were performed to completely remove all primers and other short DNA fragments. A sub-sample of each of the libraries was checked on an Agilent Tapestation, using a High Sensitivity DNA Tapescreen (Agilent) to determine the size range and concentration of each library and an equimolar pool of all libraries was made and diluted to contain ~$30 \times 10^6$ molecules per microliter.

Emulsion PCR and Ion Torrent sequencing were performed as described in Chapter 3, Material and Method Section.

### 4.2.5 Processing of sequencing data

Primer trimming

Cutadapt v1.1 (Martin, 2011) was used for the primer trimming from the raw reads of AP-based dataset using a strict zero mismatch threshold (parameters for P5: -b [Forward; GGAGGTGGTGTTCGAGGG] –b [Reverse complement; CCCTCGAACACCACCTCC] and parameters for P6: -b [Forward; GAGATTGACCTGCACGCC] –b [Reverse complement; GGCGTGCAGGTCAATCTC]). Reads with a length less than 50 nucleotides were removed during the primer trimming process (parameter: –m [50 # discard reads that are shorter than min-length]).

Quality filtering and data annotation

After primer trimming datasets were uploaded to the Metagenome Rapid Annotation using Subsystem Technology (MG-RAST) server

(http://metagenomics.nmpdr.org/) (Meyer et al. 2008), where low-quality reads and artificial replicates were removed according to MG-RAST default settings. Datasets were annotated with the protein M5NR (M5 non-redundant) database resulting in protein-derived taxonomic profiles (MG-RAST Manual v. 3.3.6, Wilke et al., 2014). The MG-RAST default annotation parameters such as maximum E-value < $1x10^{-5}$, minimum length of alignment of 15 amino acids, and minimum sequence identity of 60% were used to identify the best database matches.

### 4.2.6 Statistical metagenomic profile comparison

For comparison of the metagenomic profiles, the relative abundance scores for each taxon within a dataset were determined by the percentages of respective reads over the total annotated reads. In the text the relative abundance scores found for the taxonomic features are represented as an average ±SD (standard deviation) across all datasets. Relative abundance scores of taxonomic profiles at the phylum level are available in the Supplementary material for each of soil sample dataset (Appendix C).

To determine the difference between two taxonomic profiles, the Statistical Analysis of Metagenomic Profiles (STAMP) software package was used (Parks & Beiko 2010). Fisher's Exact Test was performed, and taxa with p-values < 0.05 were considered to be significantly different between the metagenomic profiles.

Statistical comparison of metagenomic profiles was conducted on the square root transformed data using the statistical package Primer v.6 for Windows (Version 6.1.13, PRIMER-E, Plymouth) (Clarke & Gorley 2006). All statistical approaches applied were as per described in Chapter 2, including CLUSTER, NMDS, ANOSIM,

and SIMPROF. SIMPER ('similarity percentage') analysis was used for calculation of average intergroup dissimilarity and intragroup similarity of metagenomic profiles.

Metagenome profiles were further analysed using canonical analysis of principal coordinates (CAP) using the PERMANOVA+ version 1.0.3 3 add-on to PRIMER 6 (Anderson MJ, Gorley RN 2008) as a constrained ordination method to test the significance of the differences among the *a priori* groups in multivariate space. All metagenomic profiles were divided into groups according to the origin of the samples. The *a priori* hypothesis of 'no difference' within groups was tested using CAP analysis by evaluation of a *p*-value obtained after 9999 permutations. The strength of the association between multivariate data and the hypothesis of group differences was indicated by the value of the squared canonical correlation ($\delta_1^2$). An appropriate number of principal coordinates axes (m) used for the CAP analysis were chosen automatically by the CAP routine to minimize errors of a misclassification. In order to validate the ability of the CAP model to classify correctly the samples according to their appropriate groups, a cross-validation procedure was performed for the chosen value of m. Classification of the new unknown samples into the existing groups was also performed with CAP (Anderson & Willis 2003).

*Step-by-step procedure of a likelihood ratio (LR) model computation* was as described in the Materials and Methods Section of Chapter 3.

## 4.3 Results and discussion

### 4.3.1 Soil sampling and samples notation

Four sets of samples were collected at different times of the year from each of three sites (Table. 4.2). Set 1 consisted of four replicates of soil samples from each location collected 1 m apart. Sets 2, 3, and 4 were collected at three other time points at different seasons and included only one soil sample taken from each location. In total 21 soil samples were collected and analysed.

### 4.3.2 AP-PCR based high throughput DNA sequencing for soil discrimination

The AP-PCR based high throughput sequencing method proposed in this research consists of two main stages, namely amplification of soil metagenomic DNA using the AP-PCR approach and massive parallel sequencing of the obtained amplification products. To ensure quality, reproducibility, and reliability of the entire method both stages should be rigorously assessed. This must include determining conditions for AP-PCR amplification in order to obtain reproducible and band-rich DNA patterns prior to the evaluation of reproducibility of the HTS DNA sequencing. Annotation of sequencing data and its statistical analysis also play a pivotal role in revealing an adequate picture of the soil microbial community structure.

### 4.3.3 Optimisation of AP-PCR amplification procedure

Single arbitrarily primed PCR amplification, being a PCR-based method, requires standard optimisation as such as any other PCR-based methods. The optimal

concentration of magnesium ions ($Mg^{2+}$), appropriate annealing temperature and amount of DNA template are the most common parameters to be optimised in order to obtain satisfactory amplification results. In this study, these parameters of AP-PCR amplification were selected by conducting the PCR with the primer Seq5-RC (P5), effective in previous investigations (Waters et al. 2012), and using only one soil DNA extract from sample Aw1 (Table 4.2).

Annealing temperature. Historically, AP-PCR amplification thermocycling conditions consisted of two steps (Welsh & McClelland 1990). The first step was 2 cycles at low stringency conditions (low primer annealing temperature), followed by a second step of 40 cycles at high primer annealing temperature (stringent amplification). This temperature profile was applied based on the expectation that during first two low stringency cycles a primer with an arbitrarily chosen sequence would anneal at numerous priming sites. At the same time, it was expected that hybridisation of DNA strands at the temperatures considerably lower than their characteristic melting temperatures would allow for the formation of DNA duplexes with multiple mismatched base pairs. This could therefore result in unpredictable and non-reproducible primer binding patterns. Given that after these low stringency amplification cycles all PCR products generated would have exactly the same sequence at their 5'- and 3'-termini as the primer, the subsequent high stringency amplification should have no, or minimal, contribution to the total amplification specificity and reproducibility.

For the reason outlined above, and also taking into account that previous studies showed that the RAPD profiles become more reproducible with an increase in the annealing temperature (Paraguison et al. 2012), it was decided to perform optimization

of the annealing temperature of the first two cycles of the AP-PCR. A range of temperatures from 30 to 55 $^{o}$C was tested. Figure 4.1 shows the results of amplification of 10 ng of total DNA extracted from soil sample Aw1 at each temperature tested in triplicates. It can be seen clearly that at the low temperature smeared bands dominate. These 'smeared bands' are, in fact, multiple PCR products generated from primers bound to target DNA with a few mismatches, that are created due to binding at a low annealing temperature. As soon as thermodynamic stability of these 'mismatch annealed' primers is weak therefore the mismatched primer binding disappears as the annealing temperature of the first two cycles of AP-PCR increases. When at the higher annealing temperatures, the bands representing the PCR products become more distinct from the background. The resulting gel image suggests that high stringency AP-PCR results in preferential amplification of random targets flanked by sequences fully complementary to the arbitrary primer, or with very few mismatches. Comparison of two or more amplifications from the same DNA template under the same conditions is an easy way to conduct a quick preliminary estimation of the method's reproducibility. It would be expected that the use of high annealing temperatures at the first two cycles of AP-PCR amplification generates more specific, reproducible and predictable binding (with less mismatched nucleotide pairs) of primer to its target sequence.

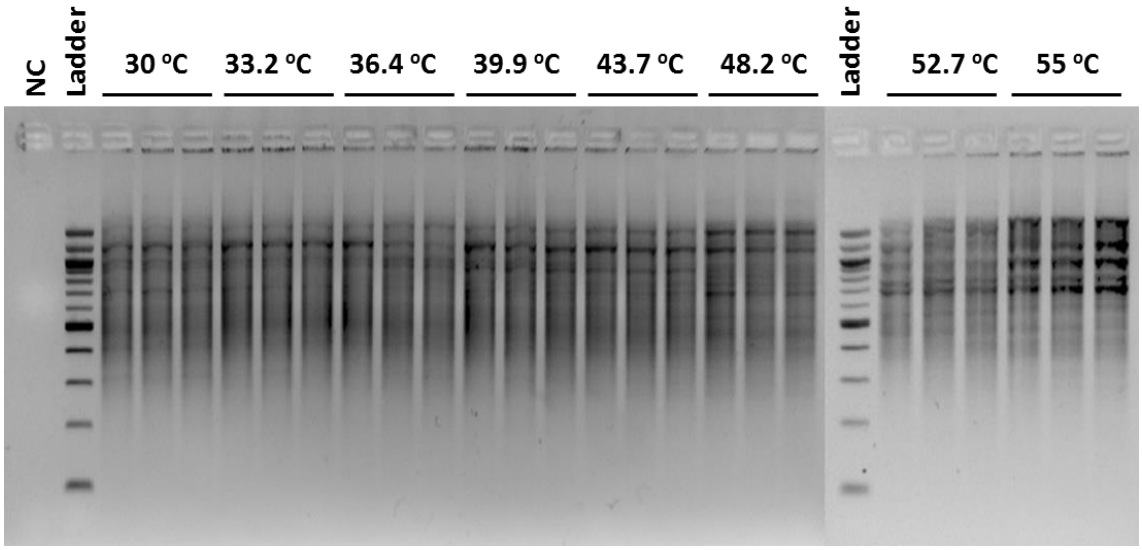



**Figure 4.1. Annealing temperature optimisation** of AP-PCR amplification with Seq5-RC primer of total DNA extracted from soil sample Aw1. NC = no template control, Ladder = Hyper Ladder II (Bioline). All reactions at different temperatures were performed in triplicates.

As a result of annealing temperature optimisation, a PCR amplification with 42 cycles consisting of denaturation at 94 $^{o}$C for 30 s, primer annealing at 55 $^{o}$C for 30 s and primer extension at 72 $^{o}$C for 1 min was selected.

Effect of $Mg^{2+}$ concentration. Varying magnesium ions concentration influences the specificity of amplification by changing the efficiency of primer hybridisation (Pelt-Verkuil et al. 2008). With a higher $Mg^{2+}$ concentration, the PCR exhibits greater tolerance to the formation of mismatched hybridisation duplexes during the annealing stage. Investigation of the $Mg^{2+}$ concentration ranging from 1.5 mM to 4 mM was performed. Figure 4.2 shows that at low magnesium ion concentration (1.5 mM), high molecular weight amplification products (in a range of 1.5 – 2 kbp) were generated. The increase in $Mg^{2+}$ up to 4 mM shifts the range of AP-PCR products being produced towards the low molecular weight region (0.5 – 1.3 kbp). This can likely be explained by a higher number of thermodynamically favourable sites available for primer annealing at these conditions. A concentration of 2.5 mM was selected as optimal as it

had produced the most intense and diverse DNA profile with band size varying in the range from 0.5 to 2 kbp.

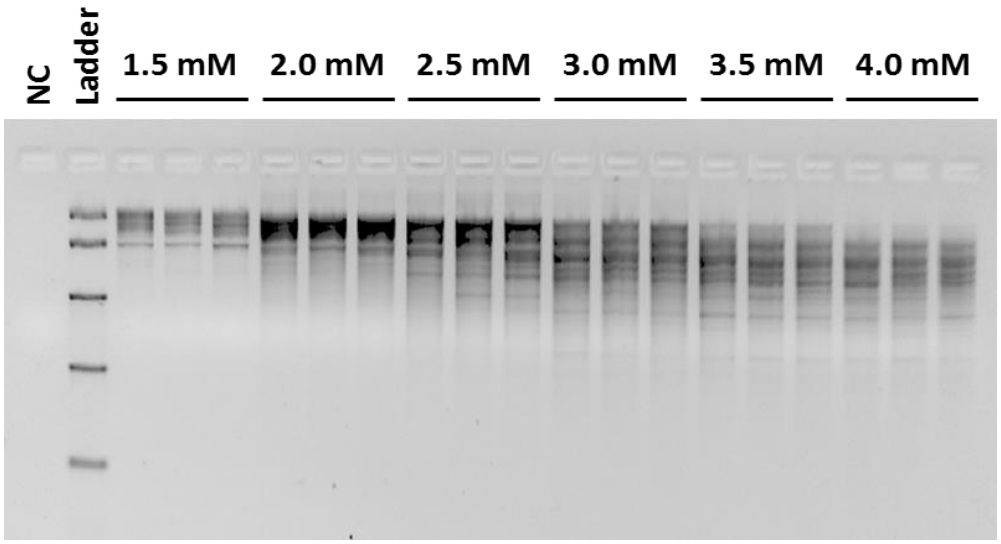



**Figure 4.2. Effect of $Mg^{2+}$ concentration** on AP-PCR amplification with Seq5-RC primer of total DNA extracted from soil sample Aw1. NC = no template control, Ladder = Hyper Ladder II (Bioline). All reactions at different temperatures were performed in triplicates.

The amount of DNA template is very important for generation of reproducible band patterns (metagenomic profiles) (Atienzar & Jha 2006). PCR is quite tolerant to the amount of high quality DNA being used as a template. Even up to 1 μg of genomic DNA can be put into the reaction with no effect on its outcome. The opposite situation is observed when DNA extracted from a soil sample is used. The more DNA that is used – the more chances to inhibit the reaction with PCR inhibitors co-extracted from the soil sample. In this study, the amount of DNA extracted from soil sample Aw1 used in the PCR was in the range from 0.125 ng to 32 ng per 25 μL. As shown in Figure 4.3, the quantity range of 1-8 ng provided a stable and rich DNA pattern. Higher amounts of DNA template (32 – 16 ng per reaction) resulted in the same band pattern but of less intensity, which can likely be explained by the presence of the minimal amount of inhibitors. Non-reproducible patterns, even within the same replicates, were observed for AP-PCR with 0.5 – 0.125 ng of DNA per reaction. These pattern-variations are

indicative of a stochastic effect of AP-PCR, especially taking into account the complexity of soil DNA assemblage used as an amplification template. For further experiments 4 ng of DNA was used.

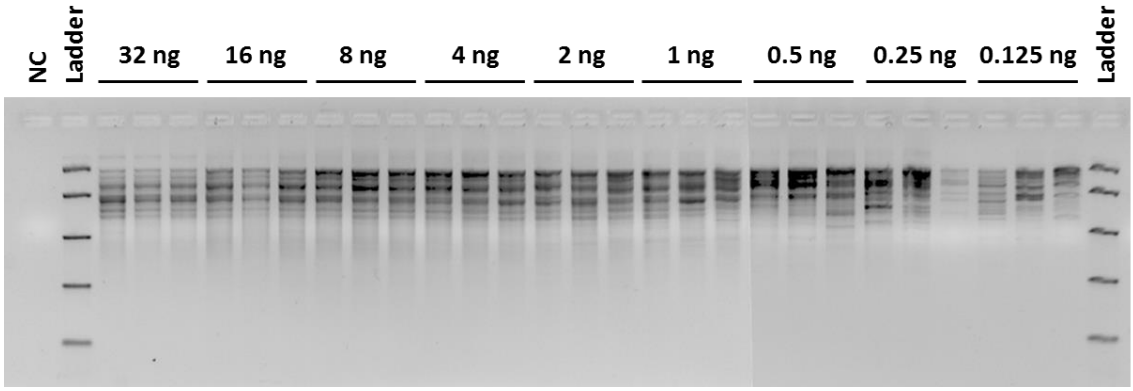



**Figure 4.3. Effect of the initial amount of DNA template used** in AP-PCR amplification with Seq5-RC primer of total DNA extracted from soil sample Aw1. NC = no template control, Ladder = Easy Ladder II (Bioline). All reactions were performed in triplicates.

Selection of arbitrary primer. Besides Seq5-RC, which was proven to be efficient at generating diverse DNA patterns during AP-PCR (Waters et al. 2012), five additional arbitrarily chosen primers (Table 4.3) were tested for their performance at the selected AP-PCR conditions. One primer represented the reverse complement sequence of the Seq5-RC primer (primer Seq5); two more primers (primer Seq6 and primer Seq7) were randomly generated using Random DNA Sequence Generator software freely available at the http://www.faculty.ucr.edu/~mmaduro/random.htm website. The following constrains were used during the design of these primers: the GC content was set at approximately 60% with a length of 18 nucleotides. The last two primers Seq6-RC and Seq7-RC were the reverse complements of the primer Seq6 and primer Seq7, respectively. AP-PCR amplification of the DNA extracted from the soil sample Aw1 was performed in triplicates for each primer.

Figure 4.4 shows that among the six primers tested, only two (primer Seq5-RC and primer Seq6) were successful at generating rich DNA profiles with multiple well-resolved bands. The rest of the primers showed only high molecular weight PCR products with significantly lower yields. Primers Seq-5RC and Seq6 were subsequently used for AP-PCR amplification of soil metagenomic extracts followed by HTS.

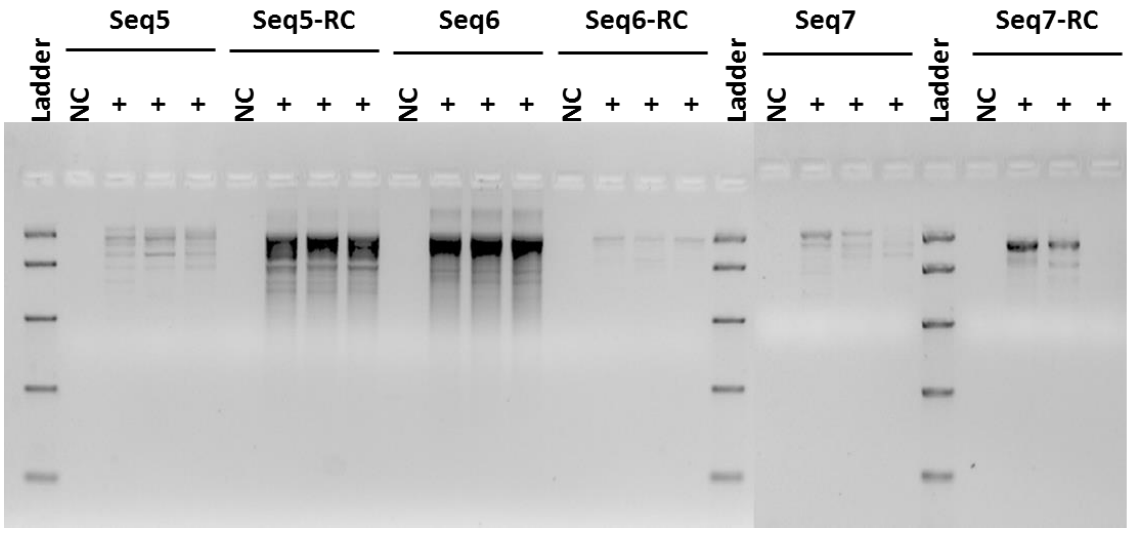



**Figure 4.4. Evaluation of six arbitrarily chosen primers for AP-PCR amplification**. As a template 4 ng of DNA extracted form soil sample Aw1 was used. NC = no template control, Ladder = Hyper Ladder II (Bioline). Each amplification reaction with different primers was performed in triplicates.

Generation of DNA profiles of three different soil types using AP-PCR amplification. Two selected arbitrarily chosen primers (Seq5-RC and Seq6) were consequently employed for generation of DNA profiles from total DNA extracted from soil samples Aw1, Bw1 and Cw1. DNA (4 ng) was amplified using selected amplification conditions of $Mg^{2+}$ concentration at 2.5 mM, annealing temperature of 55 $^{o}$C and 42 cycles. Each amplification reaction was performed in triplicates to ensure reproducibility of the selected conditions for the different soil types. Agarose electrophoretic analysis of the amplification reaction is shown in Figure 4.5. For each soil sample type, a visually different band pattern was obtained. Each pattern

represented a high molecular weight smear (0.5 – 2 kbp) with bright bands scattered throughout. The results confirm that optimised conditions for AP-amplification allow for the generation of relatively reproducible and highly specific DNA fingerprints for soil samples used in the study.

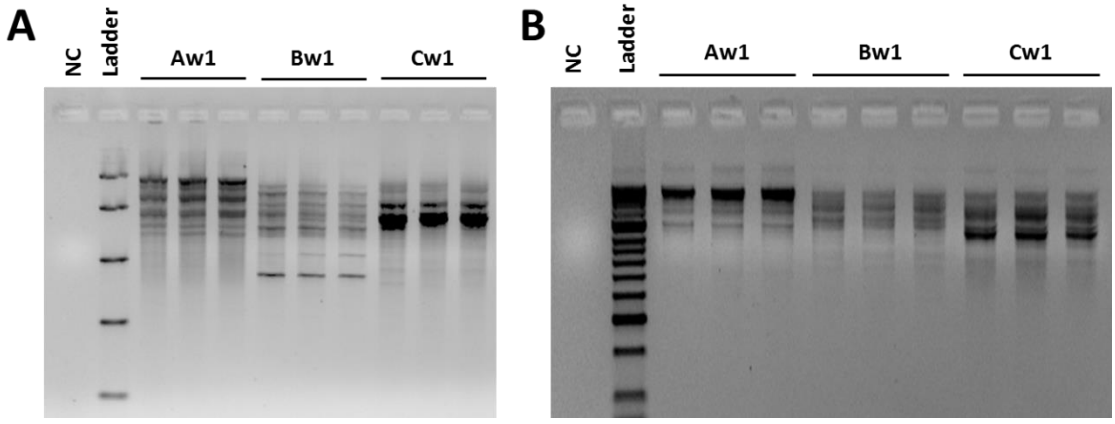



**Figure 4.5. DNA profiles (fingerprints) generated using AP-PCR technique with (A) Seq5-RC and (B) Seq6 primers** from total DNA specimens extracted from soil samples Aw1, Bw1 and Cw1. Each reaction with different DNA specimen was performed in triplicates. NC = no template control, Ladders: (**A**) Easy Ladder I (**B**) Hyper Ladder II (Bioline).

### 4.3.4 General characteristics of obtained AP-PCR based sequence datasets

Out of 21 soil samples collected (Table 4.3), 26 sequence datasets were generated using AP-PCR based sequencing with primer Seq5-RC (P5) and 12 datasets with primer Seq6 (P6) (Table 4.4).

**Table 4.4. Notation of metagenomic DNA sequence datasets.**

| Sampling sites | Set 1 | | Set 2 | Set 3 | Set 4 |
|---|---|---|---|---|---|
| | Winter (w) | | Spring (sp) | Summer (s) | Autumn (a) |
| **Flinders University (A)** | AP-Aw1′ | AP-Aw1_P6 | AP-Asp | AP-As | AP-Aa |
| | AP-Aw1′′ | | | | AP-Aa* |
| | AP-Aw1′′′ | | | | |
| | AP-Aw2 | AP-Aw2_P6 | | | |
| | AP-Aw3 | AP-Aw3_P6 | | | |
| | AP-Aw4 | AP-Aw4_P6 | | | |
| **Warradale reserve (B)** | AP-Bw1 | AP-Bw1_P6 | AP-Bsp | AP-Bs | AP-Ba |
| | AP-Bw2 | AP-Bw2_P6 | | | AP-Ba* |
| | AP-Bw3 | AP-Bw3_P6 | | | |
| | AP-Bw4 | AP-Bw4_P6 | | | |
| **Brighton Esplanade (C)** | AP-Cw1 | AP-Cw1_P6 | AP-Csp | AP-Cs | AP-Ca |
| | AP-Cw2 | AP-Cw2_P6 | | | AP-Ca* |
| | AP-Cw3 | AP-Cw3_P6 | | | |
| | AP-Cw4 | AP-Cw4_P6 | | | |

* Metagenomic DNA datasets are named as follows: all names start from a prefix "AP" which shows that the dataset was obtained employing AP-PCR amplification; the next capital letter such as 'A', 'B' or 'C' shows a location of soil sampling, namely site A, B or C, respectively; the next small letter represents season (time) of sample collection such as 'a' for autumn, 'w' for winter, 'sp' for spring and 's' for summer; an asterisk (*), if any, represents the sample sequenced at ACRF Cancer Genomic Facility; a following digit (from 1 to 4), if any, represent a number of replicates taken from the location, extracted, amplified and then sequenced; number of acute accent symbols (′), if any, represents a number of replicative AP-PCR products obtained from the same metagenomic DNA and then sequenced; 'P6', if any, shows that this particular metagenomic DNA sample has been amplified with Seq6 primer (Table 4.2), absence of 'P6' shows that the metagenomic DNA sample has been amplified with Seq5-RC primer (Table 4.2).

Sequencing of the amplification products obtained with the P5 primer resulted in an average of 272,244 (74,370 – 1,047,266) sequence reads for a total of 43.4 Mbp (8.9 – 171.1 Mbp) of sequence (Table 4.5). Sequencing datasets after AP-PCR amplification with the P6 primer consisted of an average of 192,840 (137,711 – 295,706) sequences for a total of 39.2 Mbp (28.3 – 59.5 Mbp). Primer trimming resulted in a decrease of the average number of reads by 8% and 3% for P5 and P6 based datasets respectively. The obtained datasets (after primer trimming) were uploaded to the online MG-RAST server (Meyer et al. 2008), where approximately 49% (P5) and

42% (P6) of low quality reads were eliminated from each dataset respectively. Finally 120,619 (28,061 – 457,455) (P5 primer) and 106,822 (64,197 – 156,270) (P6 primer) sequencing reads were available for subsequent annotation (Table 4.5). Predicted protein regions with known functions were found in 62% and 67% (average) of trimmed high quality reads for the P5 and P6 datasets. On average, 17% (P5) and 19% (P6) of high quality reads had predicted protein regions with unknown functions. The number of unassigned reads after trimming and quality filtering constituted 20% and 12% for the P5 and P6 datasets, respectively (Table 4.6).

**Table 4.5. General characteristics of obtained AP-PCR based sequence datasets**

| Datasets | Number of reads (range) | Primer trimming (%) | Number of reads uploaded to MG-RAST (range) | Number of Mbp | Read length, bp ± SD | GC , % ± SD | Failed QF (% of total) | Dereplication (% of total) |
|---|---|---|---|---|---|---|---|---|
| **AP-PCR (Seq5_RC)** | 272,244 (74,370 – 1,047,266) | 8 | 252,515 (66,030 – 973,419) | 43.4 | 165 ± 51 | 58 ± 6 | 49 | 31 |
| **AP-PCR (Seq6 )** | 192,840 (137,711 – 295,706) | 3 | 186,253 (133,496 – 284,338) | 39.2 | 210 ± 56 | 57 ± 5 | 42 | 29 |

SD = Standard deviation, QF= Quality Filtering

**Table 4.6. MG-RAST annotation of obtained AP-PCR based sequence datasets**

| Datasets | Number of reads passed QF | Number of reads with predicted protein regions (% of QF reads) | | | | Number of reads with predicted rRNA genes (% of QF reads) | | Number of unassigned reads (% of QF reads) | | Number of predicted protein regions (features) | Number of assigned features to M5NR database (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | known | | unknown | | | | | | | |
| **AP-PCR (Seq5_RC)** | 120,619 (28,061 – 457,455) | 79,119 (12,945 -296,117) | 62 | 16,812 (5,726 – 52,467) | 17 | 210 (0 - 955) | 0.15 | 22,058 (4,654 – 94,146) | 20 | 23,919 (6,798 – 80,330) | 20 |
| **AP-PCR (Seq6)** | 106,822 (64,197 – 156,270) | 71 647 (44,430 - 101,840) | 67 | 21,306 (8,650 – 34,727) | 19 | 977 (0 -10,543) | 0.11 | 12,551 (5,677 – 16,504) | 12 | 29,023 (11,109 – 46,265) | 26 |

QF= Quality Filtering

*Taxonomic profiling of three soil metagenomes*

Taxonomic classification of protein gene fragments showed that 85 ± 2% of annotated reads were assigned to Bacteria, 4.0 ± 2% of reads also matched to Eukaryota and 0.3% (±0.3%) to Archaea. The remaining 10 ± 2%) of reads were not assigned. Bacterial taxa (*Proteobacteria, Actinobacteria, Bacteroidetes, Planctomycetes, Acidobacteria, Verrucomicrobia*) dominated in all metagenomic datasets representing close to 75% of annotated reads. Additional phyla including *Chloroflexi, Firmicutes, Cyanobacteria* and *Chlorobi* represented less than 5% of reads. *Proteobacteri*a was the most dominant phyla across all datasets (generated with either P5 or P6 primers) with an average abundance of 43 ± 10%). However the visual comparison of the taxonomic abundance patterns from soil samples A, B and C clearly showed that the profiles obtained with the P6 primer were more similar than those generated by the P5 primer (Figure 4.6). Datasets generated by amplification with the P6 primer contained 5 times higher relative abundance of *Verrucomicrobia* and 3 times higher abundance of *Bacteroidetes* than the datasets obtained with P5 primer. AP-PCR with the the P5 primer resulted in the notably higher representation of *Gemmatimonadetes* in soil samples from site B, 7% versus 0.3% and 0.1% for sites A and C, respectively. *Deinococcus-Thermus* was found to be more abundant in soil C (1.1%) rather that in soils A (0.3%) and B (0.2%). *Plantomycetes* were considerably less abundant in replicative samples B (1.3%) than in samples A (8.5%) and C (7%). The distribution of main taxa in datasets A, B and C generated with the P6 primer was more even. Among the eukaryotic taxa, *Ascomycota* was found to be the dominant microorganism with average abundance of 2 ± 1% in all datasets obtained (Figure 4.6).

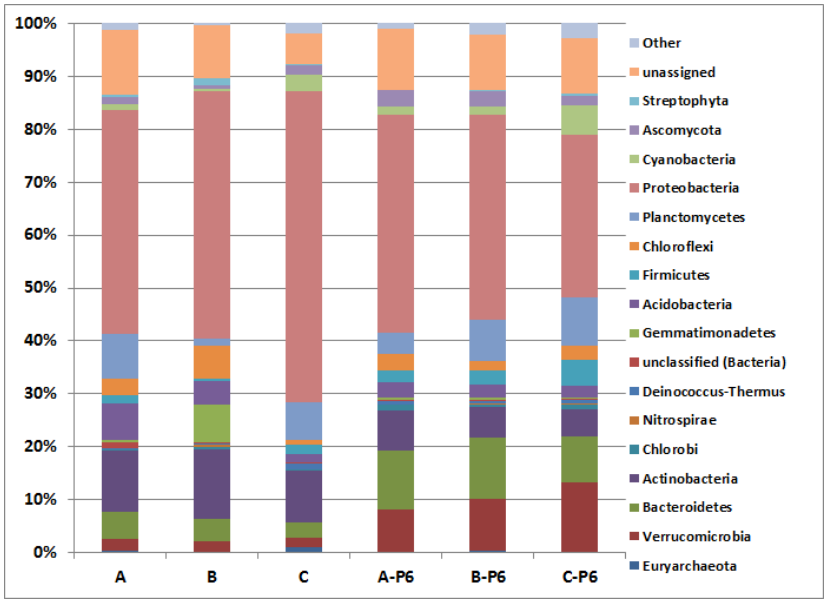



**Figure 4.6. Average relative abundances of taxa within each soil type detected by AP-PCR** amplification with primers P5 (first three columns) and P6 (last three columns).

### 4.3.5 Comparison of the taxonomic profiles and discrimination of soil samples using multivariate statistical analysis

Reproducibility of Ion Torrent library preparation and high throughput DNA sequencing procedures.

Reproducibility of the analysis is the one of the main requirements for forensic investigation. In the previous section it was shown that optimised AP-PCR amplification allows for the generation of highly reproducible gel band patterns. However it might be that many fragments in the resulting DNA fingerprint may appear identical in length but different in sequence. Thus the reproducibility of the sequence content of the AP-PCR amplification products needed to be assessed.

In order to fill this gap three replicates of AP-PCR amplification of the same DNA sample extracted from soil samples Aw1 were individually subjected to Ion Torrent library preparation and sequencing.

To perform a robust and reliable comparison of soil metagenomes, all further analyses were conducted at the highest level of taxonomic resolution (species). Taxonomic profiles of the obtained sequencing datasets (Aw1′, Aw1″ and Aw1‴) underwent statistical comparison using Fisher's Exact Test implemented in the STAMP software. This type of statistical analysis identifies features that differ significantly in their abundances. Pair-wise comparison between the taxonomic profiles of three replicates is shown in Figure 4.7. The scatter plots (Figure 4.7 A-C) demonstrate that all dots representing the individual species found in the taxonomic profiles lay on the dashed trend line with an average $R^2$ of 0.981. This indicates that the analysed profiles are highly similar. Moreover according to the Fisher's Exact Test no taxa were found with significantly different abundances between the replicate profiles.

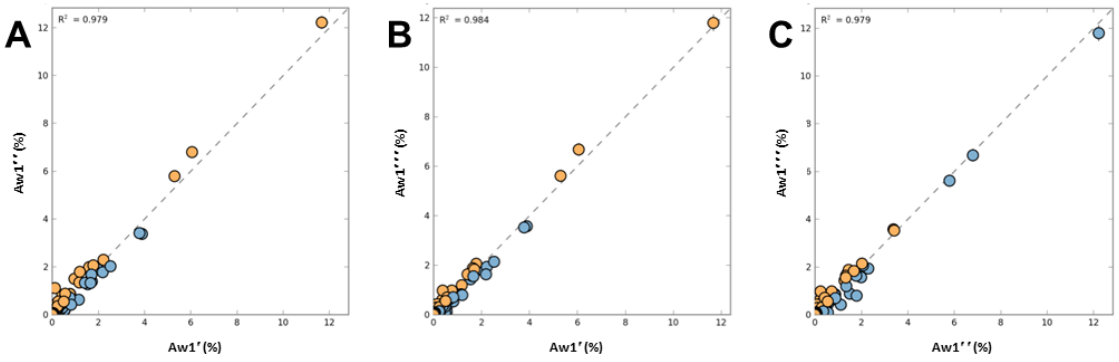



**Figure 4.7. Results of Fisher's Exact Test pair-wise comparison** of three independent replicative samples obtained by AP-PCR amplification with primers Seq5-RC extraction of total DNA from soil Aw1. **A** – pair-wise comparison of replicates AP-Aw1 and AP-Aw1″; **B** – pair-wise comparison of replicates AP-Aw1′ and AP-Aw1‴; **C** – pair-wise comparison of replicates AP-Aw1″ and AP-Aw1‴.

Variation of soil microbial composition within a site and between different sites.

Comparison of the protein-derived taxonomic profiles generated from 24 sequence datasets (3 sampling sites × 4 replicates × 2 arbitrary chosen primers) was performed using multivariate statistical analysis.

Datasets generated with Seq5-RC (P5) primer. CLUSTER analysis with group-average linking based on Bray-Curtis profile pair-wise similarity scores, summarised in Table 4.7, resulted in delineation of three distinct clusters (Figure 4.8A). Further overlaying of the obtained clusters on a NMDS plot (Figure 4.8B) was consistent with the grouping of the samples observed on the CLUSTER dendrogram, where all samples grouped according to their collection sites (Figure 4.8).

**Table 4.7. Bray-Curtis pair-wise similarity scores obtained for datasets generated with primer P5.**

| Dataset name | APAw1 | AP_Aw2 | AP_Aw3 | AP_Aw4 | AP_Bw1 | AP_Bw2 | AP_Bw3 | AP_Bw4 | AP_Cw1 | AP_Cw2 | AP_Cw3 | AP_Cw4 | AP_Aw1_P6 | AP_Aw2_P6 | AP_Aw3_P6 | AP_Aw4_P6 | AP_Bw1_P6 | AP_Bw2_P6 | AP_Bw3_P6 | AP_Bw4_P6 | AP_Cw1_P6 | AP_Cw2_P6 | AP_Cw3_P6 | AP_Cw4_P6 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **AP_Aw1** | | | | | | | | | | | | | | | | | | | | | | | | |
| **AP_Aw2** | 68 | | | | | | | | | | | | | | | | | | | | | | | |
| **AP_Aw3** | 71 | 70 | | | | | | | | | | | | | | | | | | | | | | |
| **AP_Aw4** | 72 | 68 | 71 | | | | | | | | | | | | | | | | | | | | | |
| **AP_Bw1** | 54 | 56 | 61 | 53 | | | | | | | | | | | | | | | | | | | | |
| **AP_Bw2** | 53 | 51 | 56 | 50 | 71 | | | | | | | | | | | | | | | | | | | |
| **AP_Bw3** | 57 | 58 | 63 | 56 | 68 | 66 | | | | | | | | | | | | | | | | | | |
| **AP_Bw4** | 56 | 56 | 61 | 55 | 68 | 67 | 71 | | | | | | | | | | | | | | | | | |
| **AP_Cw1** | 58 | 57 | 57 | 57 | 48 | 48 | 51 | 49 | | | | | | | | | | | | | | | | |
| **AP_Cw2** | 54 | 53 | 55 | 53 | 46 | 45 | 49 | 47 | 60 | | | | | | | | | | | | | | | |
| **AP_Cw3** | 58 | 57 | 58 | 59 | 47 | 46 | 50 | 48 | 70 | 60 | | | | | | | | | | | | | | |
| **AP_Cw4** | 55 | 52 | 54 | 54 | 45 | 43 | 47 | 46 | 60 | 55 | 58 | | | | | | | | | | | | | |
| **AP_Aw1_P6** | 55 | 57 | 56 | 58 | 47 | 46 | 51 | 48 | 53 | 51 | 54 | 49 | | | | | | | | | | | | |
| **AP_Aw2_P6** | 56 | 58 | 58 | 58 | 46 | 44 | 50 | 49 | 52 | 51 | 54 | 50 | 66 | | | | | | | | | | | |
| **AP_Aw3_P6** | 53 | 58 | 55 | 56 | 47 | 45 | 50 | 48 | 51 | 50 | 53 | 48 | 63 | 63 | | | | | | | | | | |
| **AP_Aw4_P6** | 56 | 56 | 58 | 59 | 46 | 44 | 50 | 48 | 52 | 51 | 54 | 49 | 66 | 65 | 59 | | | | | | | | | |
| **AP_Bw1_P6** | 56 | 59 | 59 | 60 | 49 | 45 | 52 | 49 | 53 | 52 | 55 | 51 | 63 | 65 | 63 | 63 | | | | | | | | |
| **AP_Bw2_P6** | 56 | 60 | 60 | 60 | 48 | 45 | 52 | 49 | 54 | 52 | 56 | 51 | 64 | 65 | 64 | 62 | 76 | | | | | | | |
| **AP_Bw3_P6** | 56 | 58 | 58 | 59 | 47 | 44 | 50 | 48 | 52 | 53 | 54 | 50 | 63 | 63 | 63 | 64 | 74 | 74 | | | | | | |
| **AP_Bw4_P6** | 56 | 59 | 59 | 59 | 48 | 44 | 51 | 49 | 54 | 52 | 56 | 51 | 65 | 66 | 66 | 64 | 76 | 77 | 74 | | | | | |
| **AP_Cw1_P6** | 56 | 57 | 58 | 59 | 48 | 44 | 50 | 49 | 54 | 53 | 55 | 51 | 62 | 62 | 61 | 62 | 65 | 67 | 67 | 66 | | | | |
| **AP_Cw2_P6** | 55 | 55 | 57 | 58 | 48 | 44 | 50 | 48 | 53 | 53 | 55 | 51 | 60 | 60 | 59 | 61 | 64 | 64 | 66 | 64 | 73 | | | |
| **AP_Cw3_P6** | 55 | 57 | 57 | 58 | 47 | 44 | 50 | 48 | 54 | 53 | 56 | 51 | 61 | 61 | 60 | 61 | 65 | 65 | 66 | 64 | 75 | 73 | | |
| **AP_Cw4_P6** | 55 | 56 | 57 | 58 | 46 | 43 | 49 | 48 | 55 | 53 | 56 | 51 | 61 | 60 | 59 | 62 | 65 | 66 | 66 | 65 | 74 | 72 | 74 | |

The results of the AP-PCR based soil microbial profiles comparison demonstrated that soil samples from the same sampling site were more similar to each other than to the samples collected from any of the other two sites. An average similarity of the profiles originating from soil samples taken within a particular location was assessed by SIMPER analysis and was found to be 70%, 68% and 61% for sites A, B and C, respectively (Table 4.8). An average pair-wise dissimilarity between groups of samples collected from location A compared to those from location B or C was 44%, whereas between the groups of samples from location B and location C it was 53% (Table 4.8). To test the significance of the identified differences between the groups ANOSIM analysis was then conducted. This showed that the metagenomic profiles originating from the samples collected from locations A, B and C were significantly different between each other (Global R = 0.927; $p < 0.0002$).

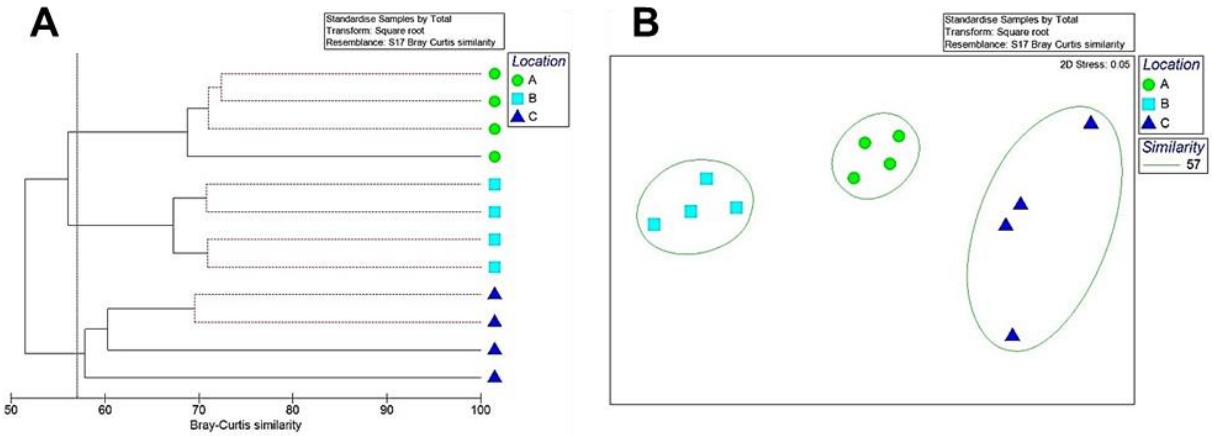



**Figure 4.8. Comparison of the taxonomic profiles (species level) generated from three soil types (A, B and C) by AP-PCR based sequencing with primer P5**. Bray-Curtis distance similarity matrix was calculated from the square-root transformed abundance of DNA fragments matching taxa in the M5NR database (E-value < $1\times10^{-5}$). The Bray-Curtis matrix was used for generating CLUSTER dendrogram and NMDS ordination plot. (**A**) Red dashed branches on the CLUSTER dendrogram indicate no significant difference between metagenomic profiles (supported by the SIMPROF analysis, $p<0.05$). (**B**) The NMDS ordination plot displays distances between samples. Data points that are closer to each other represent samples with higher similarity of metagenomic profiles. A contour line on the NMDS plot drawn round each of the clusters defines the superimposition of clusters from CLUSTER dendrogram at the selected level of similarity.

Performance of the AP-PCR with primer Seq6 (P6) for soil DNA profiling followed by HTS was evaluated in the same way as for primer P5. CLUSTER and NMDS ordination analyses demonstrated the formation of three separate clusters consisting of taxonomic profiles grouped correctly according to their origin (Figure 4.9). It can be seen that four samples from locations B and C fuse into two separate clusters, having an average Bray-Curtis profile similarity of 75% and 73%, respectively.

**Table 4.8. SIMPER analysis of an average intra-group profiles similarity and an average inter-group dissimilarity.**

| **Average *similarity* within site, %** | | |
|---|---|---|
| | ***Primer Seq5-RC*** | ***Primer Seq6*** |
| Group A | **70** | **64** |
| Group B | **68** | **75** |
| Group C | **61** | **73** |
| **Average *dissimilarity* between sites, %** | | |
| | ***Primer Seq5-RC*** | ***Primer Seq6*** |
| Groups A & B | **44** | **36** |
| Groups A & C | **44** | **39** |
| Groups B & C | **53** | **35** |

Samples from location A formed a genuine cluster consisting only of three soil metagenomic profiles. The profile from the fourth replicate sample of soil A formed a separated branch on the dendrogram closely related to the cluster of the A samples (Figure 4.9). An average similarity between all profiles from site A was 64% according to SIMPER analysis. The average dissimilarity between soil profiles A and B was found to be 36%, between A and C 39%, and between B and C 35%. ANOSIM analysis confirmed the significance of the differences found between these groups of samples ($R=0.711$, $P<0.0002$).

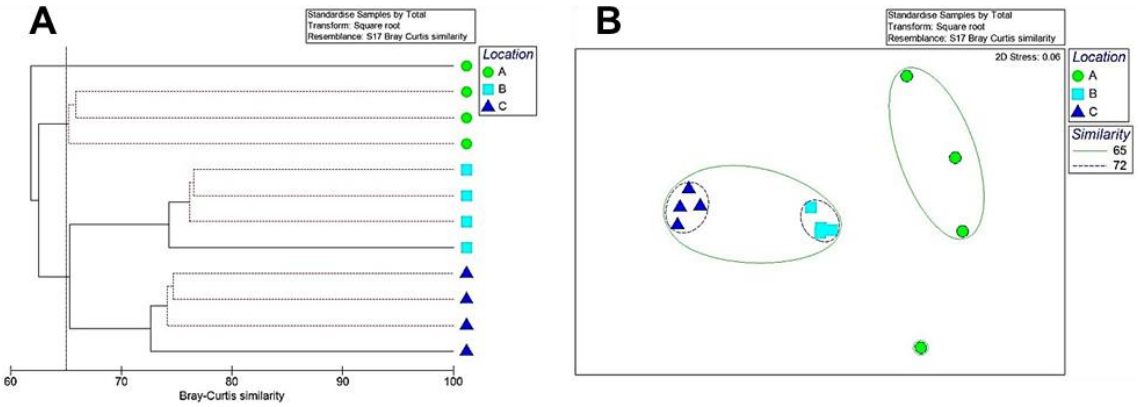
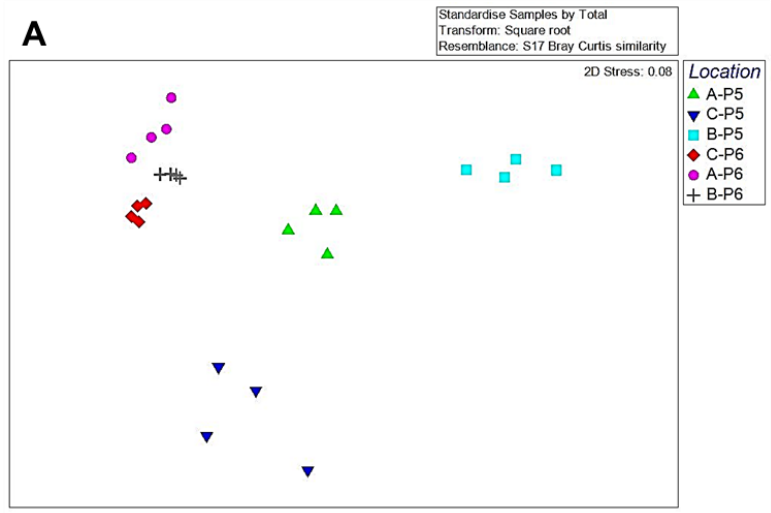



**Figure 4.9. Comparison of the taxonomic profiles (species level) generated from three soil types (A, B and C) by AP-PCR based sequencing with primer P6**. Bray-Curtis distance similarity matrix was calculated from the square-root transformed abundance of DNA fragments matching taxa in the M5NR database (E-value < $1\times10^{-5}$). The Bray-Curtis matrix was used for generating CLUSTER dendrogram and NMDS ordination plot. (**A**) Red dashed branches on the CLUSTER dendrogram indicate no significant difference between metagenomic profiles (supported by the SIMPROF analysis, $p<0.05$). (**B**) The NMDS ordination plot displays distances between samples. Data points that are closer to each other represent samples with higher similarity of metagenomic profiles. A contour line on the NMDS plot drawn round each of the clusters defines the superimposition of clusters from CLUSTER dendrogram at the selected level of similarity.

Simultaneous comparison of the taxonomic profiles generated by primer P5 and primer P6 from the same soil samples resulted in the formation of six clearly separated groups of samples on the NMDS plot (Figure 4.10A). Clusters formed by P6-based taxonomic profiles positioned closer to each other, indicating their higher similarity than the clusters formed by P5-based profiles. Results of the ANOSIM analysis also confirmed the significance of the observed soil samples separation (Global R = 0.891, $p<0.0001$). It is interesting to note that the high pairwise R values were observed when comparing groups formed by the profiles generated from the same soils but using different primers (P5 and P6 groups from site A: R= 0.99; site B: R=1 and site C: R=0.979) (Figure 4.10B). This probably indicates that single arbitrarily chosen primers select/amplify DNA from different taxa present in the same metagenome, which in turn results in the generation of significantly different taxonomic profiles of the same soil samples.




B

| Pair-wise Tess Groups | R Statistic |
|---|---|
| A-P5, C-P5 | 0.901 |
| A-P5, B-P5 | 1 |
| A-P5, C-P6 | 1 |
| A-P5, A-P6 | 0.99 |
| A-P5, B-P6 | 1 |
| C-P5, B-P5 | 1 |
| C-P5, C-P6 | 0.979 |
| C-P5, A-P6 | 1 |
| C-P5, B-P6 | 0.979 |
| B-P5, C-P6 | 1 |
| B-P5, A-P6 | 1 |
| B-P5, B-P6 | 1 |
| C-P6, A-P6 | 0.833 |
| C-P6, B-P6 | 1 |
| A-P6, B-P6 | 0.563 |

**Figure 4.10. Comparison of the taxonomic profiles (species level) generated from three soil types (A, B and C) by AP-PCR based sequencing with primer P5 and primer P6**. (**A**) Bray-Curtis distance similarity matrix was calculated from the square-root transformed abundance of DNA fragments matching taxa in the M5NR database (E-value < $1\times10^{-5}$). The Bray-Curtis matrix was used for generating NMDS ordination plot. The NMDS plot displays distances between samples. Data points that are closer to each other represent samples with higher similarity of metagenomic profiles. (**B**) ANOSIM analysis of profiles differences between groups of samples separated by the combined factor: primer used for soil DNA amplification and origin of soil.

The obtained results demonstrated clearly that AP-PCR based sequencing has considerable potential to discriminate between forensic soil samples from different sites. The high similarity of the profiles from the samples taken within a site rather than between spatially separated locations suggest that the proposed approach was able to reproducibly detect the unique biological signal within a small scale (1 $m^2$) of the particular area. These findings hold great promises for forensic investigations because the collection of control, reference or alibi soil samples might occur from slightly different spaces within a target location. In addition an important result of the study is that visually similar soil samples, such as from sites A and B of similar land use and similar vegetation type, were successfully discriminated. This would otherwise represent a challenge for forensic investigators.

The principal concept of the application of AP-PCR based metagenomic sequencing for the generation of unique and reproducible soil metagenomic DNA

profiles was successfully tested using two different primers. Each of these primers allowed for site specific discrimination of three soils taken from different urban areas, and what is more important, two of these soil samples were visually very similar. It was also shown that different nucleotide composition of the primer used for AP-PCR amplification allowed for selection and amplification of significantly different parts of the soil metagenome. Further investigation of this phenomenon is of much interest regarding the ability of the method to amplify and identify rare/low abundant species present in soils.

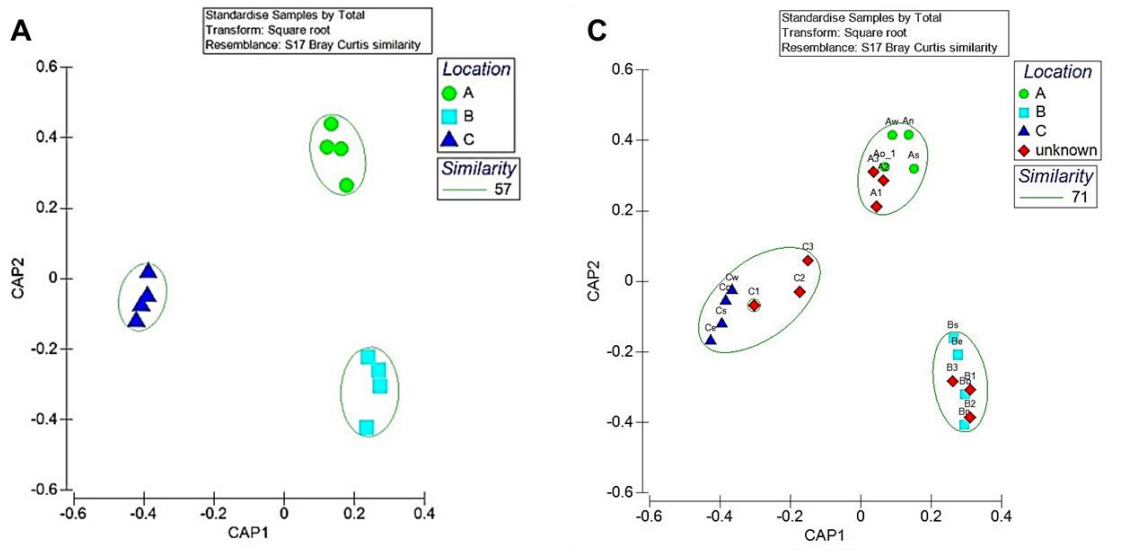
Seasonal variability of the soil metagenomes.

Examining the temporal variability in the soil microbial community has a critical importance for forensic soil analysis. The collection of soil samples from a crime scene or other control sites is often carried out with some delay from several days or even months or years after the offence has occurred (Meyers & Foran 2008). Till now there is no agreement on how much the soil microbial community changes over time. A recent study showed that commonly occurring members of the soil microbial communities exhibit minimal temporal changes while the rare taxa can undergo large changes in abundance over time (Shade et al. 2014). Some DNA-based studies indicate that spatial variability may exceed temporal variability across broad geographical gradients (Lauber et al. 2013). It was shown that temporal changes in soil microbial community structure can vary in the scale of months (Lauber et al. 2013), seasons (Voříšková & Brabcová 2014) and years (Buchan et al. 2010). An understanding of how these changes in soil biota over time will influence reliability, accuracy, reproducibility and outcome of the soil DNA comparison and discrimination is of paramount importance for forensic science.

An AP-PCR approach has primer dependant pre-selection and pre-enrichment mechanisms for metagenomic DNA amplification that emphasises the differences between samples from similar land use and vegetation type, as shown in the current study.

To assess the influence of seasonal changes on the ability of the proposed AP-PCR based sequencing technique to discriminate geographically distant soils, samples were collected from the same sites at three additional time points (Table 4.2). In combination with the previous set of samples (Set 1), new samples cover each of the seasons during the year. Total soil DNA from all collected sets of samples (Table 4.2) was extracted, amplified with the primer P5 and sequenced. Comparison of taxonomic profiles of soil samples collected at different times of the year was performed using constrained ordination analysis CAP. A very useful feature of CAP analysis is the ability to allocate new observation (samples) to already existing groups based only on their resemblances with prior observations.

The first step of the analysis was to create and validate a CAP model based on already existing profiles for subsequent classification of new observations. Three separated clusters divided by the collection site factor for samples from Set 1 are clearly seen on the CAP ordination plot (Figure 4.11A), confirming the results of the previous CLUSTER and NMDS analyses. CAP ordination of the soils' profiles at the species level demonstrated that the first squared canonical correlation was very large ($\delta_1^2$= 0.993, p=0.0007) indicating the significance of the CAP model. The cross-validation results of the CAP model confirmed the 100% correct classification of the metagenomic profiles (Figure 4.11B). The CAP routine was then used to predict to which of these three groups new samples belonged. The prediction is made on similarity of their profiles to those of existing ones.




**B**

| Original Group | A | B | C |
|---|---|---|---|
| Taxonomy level | ***species*** ($m = 2$, $\delta_1^2 = 0.993$, P = 0.0007) | | |
| % correct | 100 | 100 | 100 |
| correct/total | 4/4 | 4/4 | 4/4 |
| Misclassified to group | n/a | n/a | n/a |

**Figure 4.11. CAP discrimination of taxonomic profiles (species level) generated from three soil types (A, B and C) by AP-PCR based sequencing with primer P5.** Bray-Curtis distance similarity matrix was calculated from the square-root transformed abundance of DNA fragments matching taxa in the M5NR database (E-value < 1×10-5). The Bray-Curtis matrix was used for generating CAP ordination plots. (A) CAP analysis tests for differences among the pre-defined groups of samples in multivariate space. The significance of group separation along the canonical axis is indicated by the large value of the squared canonical correlation (δ12) and P-value. (B) Results of the cross-validation of the CAP model. (C) Classification of new profiles obtained from soil samples collected at three different times of the year from the sites A, B and C.

All new soil samples from different seasons were allocated correctly into corresponding groups according to their collection sites (Figure 4.11C). The obtained results suggest that the use of AP-PCR based sequencing generates reproducible signals from the same location throughout the year. It also indicates that the soil metagenome is stable not only during the year but that spatial variation within soils has a larger effect than temporal variation. This was in accordance with the previous investigations of soil spatial and temporal variation (Lauber et al. 2013; Berg 2013). Although these preliminary results are promising, more in-depth investigation of the ability of AP-PCR

based sequencing to discriminate soils temporally requires additional and more frequent soil sampling and analysis.

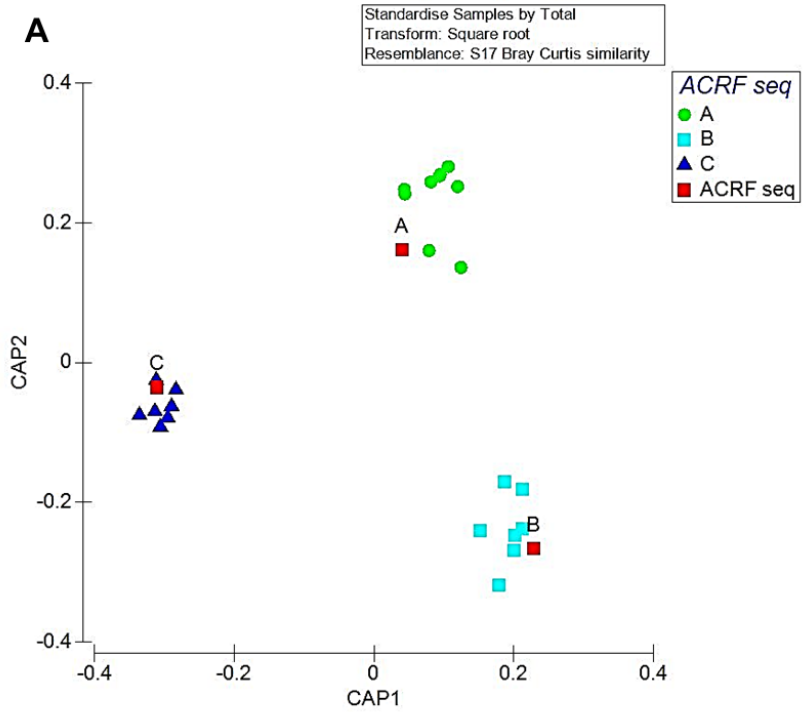
Inter-laboratory variation of high throughput DNA sequencing

To assess inter-laboratory reproducibility of the metagenomic library preparation and Ion Torrent sequencing procedure, three samples of AP-PCR products (AP-Aa, AP-Ba and AP-Ca, set 4, Table 4.4) sequenced at the Australian Genome Research Facility (AGRF) were processed in parallel at the ACRF Cancer Genomics Facility (datasets AP-Aa*, AP-Ba* and AP-Ca*, set 4, Table 4.4).

Given that all previous datasets 1-4 (Table 4.4) were successfully separated and classified into three groups according their sampling sites by CAP (Figure 4.11C), it was decided to confirm the provenance of samples Aa*, Ba* and Ca* by adding to the existing dataset groups. CAP ordination plot (Figure 4.12A) and CAP statistics (Figure 4.12B) clearly demonstrated the correct allocation of the three ACRF sequenced samples into the groups according to soil sampling sites.

Comparison of the obtained profiles from one site but generated at two different genomic facilities showed a good correlation for each pair using Fisher's Exact Test (Figure 4.13). For pairs AP-Aa/AP-Aa* and AP-Ba/AP-Ba* the correlation coefficients ($R^2$) were 0.743 and 0.728 (Figure 4.13A, Figure 4.13B, respectively). For soil sample C the $R^2$ value was 0.913 (Figure 4.13C). The average $R^2$ values of Fisher's Exact Tests for inter-laboratory reproducibility was lower compared with the one obtained for intra-laboratory reproducibility. This is likely to be explained by a difference in library preparation procedures used at the ACRF and AGRF facilities. For example metagenomic libraries at the ACRF facility were prepared using Covaris DNA shearing

procedure (http://covarisinc.com/) compared to enzymatic DNA shearing performed at AGRF.




**B**

| Original Group | A | B | C |
|---|---|---|---|
| Taxonomy level | ***species*** ($m$ = 5, $\delta_1^2$ = 0.988, P = 0.0001) | | |
| % correct | 100 | 100 | 100 |
| correct/total | 9/9 | 7/7 | 7/7 |
| Misclassified to group | n/a | n/a | n/a |

**Figure 4.12. CAP discrimination of taxonomic profiles (species level) generated at the ACRF facility from three soil samples (Aa, Ba and Ca).** Bray-Curtis distance similarity matrix was calculated from the square-root transformed abundance of DNA fragments matching taxa in the M5NR database (E-value < 1×10-5). The Bray-Curtis matrix was used for generating CAP ordination plots. (A) Classification of new profiles of soils taken from the sites A, B and C. The significance of group separation along the canonical axis is indicated by the large value of the squared canonical correlation (δ12) and P-value. (B) Results of the cross-validation of the CAP model.

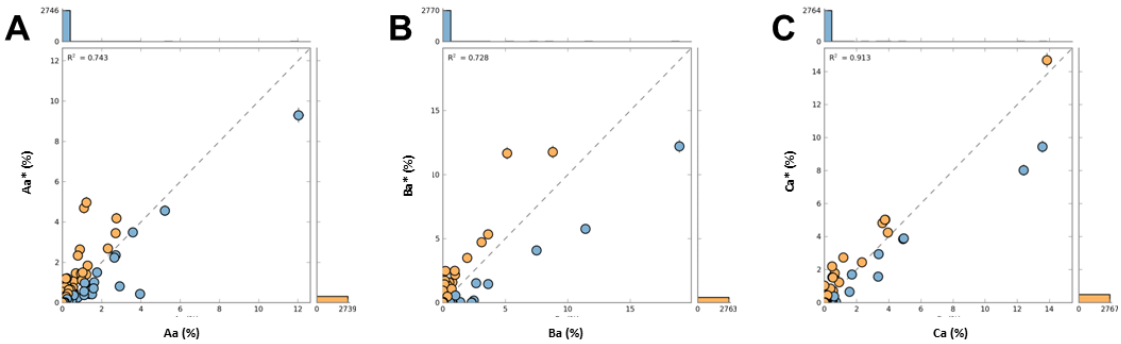



**Figure 4.13. Results of Fisher's Exact Test pair-wise comparison of the same samples proceeds at the ACRF and AGRF facilities.** Sub-samples of AP-PCR products were obtained from three soil samples (Aa, Ba and Ca) with primer P5 and subjected for library preparation and sequencing at two independent HTS providers. **A** – pair-wise comparison of replicates AP-Aa and AP-Aa*; **B** – pair-wise comparison of replicates AP-Ba and AP-Ba*; **C** – pair-wise comparison of replicates AP-Ca and AP-Ca*. '*' - means samples processed at the ACRF facility.

Characterisation of the sequencing datasets **AP**-Aa*, **AP**-Ba* and **AP**-Ca* generated at the ACRF and subsequently annotated by MG-RAST using the SEED reference database is presented in the following publication (Khodakova et al. 2013), see Appendix B.

### 4.3.6 Discriminating power of AP-PCR-based sequencing

The assessment of discriminating power and false positive / false negative error rates was carried out using the Likelihood ratio modelling proposed and described in detail in Chapter 3. For LR modeling the P5 primer AP-based sequencing metagenomic datasets obtained at AGRF facility were used (Table 4.4) which corresponded to 21 soil samples collected (Table 4.2). All major steps for the transforming of the pair-wise Bray-Curtis similarity scores between datasets into Likelihood ratio (LR) values are presented in Appendix B (Table B1, Figures B1 and B2).

The linear plot depicted in Figure 4.14 shows the correlation between similarity scores ($x$) and likelihood ratio values (in $\log_{10}$ coordinates). It can be seen that the LR line intercepts $y = 0$ ($\text{Log}_{10}(\text{LR}) = 0$) at $x = 58\%$. This indicates that the probability of obtaining similarity score above 58% is more likely if the compared samples have the same origin than different origins.

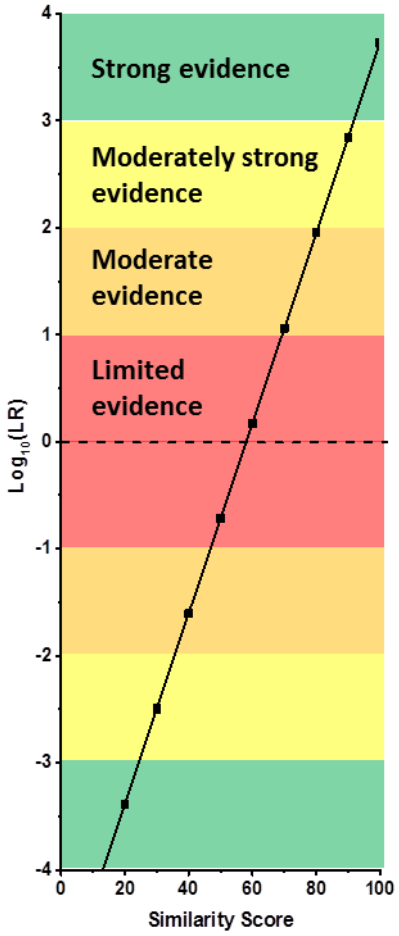



**Figure 4.14. Correlation of the decimal logarithm of LR ratios versus Similarity Scores obtained by comparing soil metagenomic profiles produced by AP-PCR-based sequencing.**

Discriminating power of the AP-based sequencing approach was further assesed on the datasets from the Set 1 (Table 4.4) generated with P5 primer. These datasets were generated from the same soil sample as those used in the Chapter 3 to enable correct comparison of the sequencing approaches performance. $\text{Log}_{10}(\text{LR})$ values

corresponding to similarity scores obtained from the comparison of visually similar soils and different soils sets in the current study were determined using a correlation plot (Figure 4.14). Within each set of $Log_{10}$(LR) values for within site and between site similarity scores were assessed Table B2 (Appendix B). Subsequently probability density functions (pdf) of the obtained $Log_{10}$(LR) values were represented on probability distribution plots with respect to the hypotheses of soil sample origin (Appendix B: $H_0$ and $H_1$ for visually similar soils (Figure B3A ); $H_0$' and $H_1$' for contrasting soils (Figure B3B)). Tippett plot curves were built to visualise false positive/ negative rates and evaluate the discriminating power of the AP-PCR based sequencing approach (Figure 4.15).

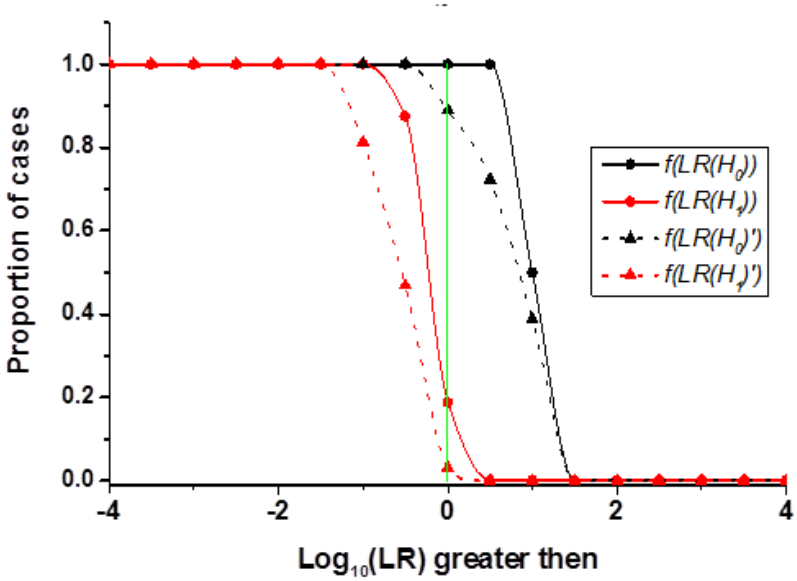



**Figure 4.15. Tippett Plot of the set of $Log_{10}$(LR) values obtained for the experimental data.** The LR is defined as a ratio of the probabilities of the evidence = similarity score (*x*) given each of two competing hypotheses $H_0$ (the pair of soils that produced score *x* originate from the same site) and $H_1$ (the pair of soils that produced score *x* originate from different sites) $H_0$ and $H_1$ for visually similar soils; $H_0$' and $H_1$' for contrasting soils.

Thus discrimination of visually similar soils showed a 0% false negative error rate, whereas the false positive error rate was estimated at a level of 19% (Figure 4.15, solid lines). Visually different soils gave an 11% false negative and 3% false positive error rate (Figure 4.15, dashed line). Overall the discrimination of both visually similar

and different soils from separate sites by an AP-PCR based sequencing approach is more relible due to lower error rates obtained than by the 16S sequecing approach explored in Chapter 3.

## 4.4 Conclusion

Soil can often be found on items collected from a crime scene and therefore can be a valuable source of information helping to link suspects to a crime scene or particular geographical location. The most challenging aspect of forensic soil testing is the discrimination of visually similar soils found at a crime scene and one found in the possession of a suspect or victim or from control sites. According to the data presented in Chapter 3 the 16S rRNA targeted sequencing method, which is a common method of microbial community analysis, showed weak discriminating power for soil samples that appear very similar or identical to the naked eye. In this chapter single arbitrary primed PCR amplification of soil metagenomic DNA followed by HTS analysis was used successfully for the discrimination of the same three soils from different urban sites as per Chapter 3. Two of these soils originated from similar land use and vegetation type sites and were visually undistinguishable. For the comparison of both visually similar and contrasting soil samples the false negative and false positive rates were shown to be considerably lower than those obtained for 16S rRNA based sequencing (evaluated in Chapter 3). That in turn shows great promises that AP-PCR based sequencing can be accepted in legal system.

According to Sensabaugh (Sensabaugh 2009) to achieve acceptance for forensic application the proposed approach of soil DNA analysis must satisfy three broad and interconnected conditions:

"1. It must be demonstrated that microbial population assemblages vary in such a way as to allow samples from a particular patch to be different from samples deriving from other places.

2. Analytical approaches for microbial profiling must be developed that combine discrimination power, robustness and reliability.

3. Statistical methods must be identified that provide objective measures for assessing the similarities and differences between samples".

The primary attempt to address those requirements for forensic general acceptance of the method has been successfully performed within the current study. This included the demonstration of different methods of multivariate statistical analysis for the measurement of similarities and differences between the soil metagenomic profiles. Evaluation of the reproducibility of the AP-PCR amplification stage and the following library preparation and sequencing stage, either at the same or different laboratories, demonstrated the reliability of the proposed approach of soil metagenomic DNA typing. Final examination of the effect of soil seasonal changes on the ability of AP-PCR based sequencing to discriminate soils revealed that the effect of temporal variations in soil microbial structure are less than differences on the soils' composition due to their collection site. Overall, the proposed combination of AP-PCR soil microbial profiling and HTS analysis in this current proof-of-concept study obtained reproducible site-specific soil microbial DNA profiles even at different times of the year.

Additionally a statistical model of soil analyses based on a Bayesian framework has been applied. In order to match the Daubert criteria of admissibility a preliminary assessment of the false negatives and false positives was performed. It should be noted that, by expanding the number and diversity of sampling sites it would be possible to refine the discrimination power of the proposed soil DNA typing method. Also the AP-PCR based sequencing method should be compared with other whole random metagenomic approaches of soil DNA typing such as shotgun and WGA sequencing.

# Chapter 5. Random whole metagenomics as a tool for forensic soil discrimination

A part of the study presented in this chapter is published as:

A.S. Khodakova, R.J. Smith, L. Burgoyne, D. Abarno, A. Linacre. Random Whole Metagenomic Sequencing for Forensic Discrimination of Soils. 2014, PLoS ONE, 9(8): e104996. doi:10.1371/journal.pone.0104996. (Appendix C).

## 5.1 Introduction

The vast majority of samples submitted for forensic investigation come from urban areas including gardens, parklands and other open spaces in built-up areas. The discrimination of geographically distinct urban soils with similar land management type and similar plant cover is of great forensic relevance (Morrisson et al. 2009; Macdonald et al. 2011). If two soil samples appear very different visually then a simple exclusion can be made but more typically soils appear visually similar and currently no further action then taken.

Development of new platforms for high-throughput DNA sequencing (HTS) has made it more affordable and led to the significant growth of HTS-based studies (Shokralla et al. 2012; Loman et al. 2012; Logares et al. 2012) including its application for forensic human DNA analysis (Fordyce et al. 2015; Daniel et al. 2015). The application of HTS to soil science has allowed for new insight into the diversity of soil microbial communities inhabiting various biomes (Fierer et al. 2012; Xu et al. 2014; Wang et al. 2013). Numerous ecological studies have shown that soil microbial communities differ between environmental habitats with different land use and vegetation type (Shange et al. 2013; Uroz et al. 2013; Fierer et al. 2012; Lauber et al. 2013; Xu et al. 2014).

Essentially there are two categories of metagenomic analysis depending on whether some specific genetic markers or whether the whole genetic assemblage is subjected to testing (Suenaga 2012). Thus in Chapter 3, 16S rRNA gene sequencing was shown to be successful at differentiating soil metagenomes derived from contrasting soil samples; however, the method demonstrated poor discrimination for visually similar soils. In Chapter 4, AP-PCR based high throughput sequencing approach was presented,

which employs a single arbitrary chosen primer for amplification of soil DNA material present in the sample. AP-PCR utilises primer-dependant sequence-specific selection of gene fragments and therefore unlikely to amplify all the DNA present in samples. Amplification with a single arbitrary primer yields an arbitrary pattern which might possess PCR products from both abundant species and those that are rare, again depending on the affinity of the primer. Also in Chapter 4, it was shown that the AP-PCR method allowed for reliable, statistically significant discriminating of both visually similar and different soil samples. The AP-PCR based sequencing approach, therefore, might be considered as PCR based method for comparative random whole metagenomic analysis.

Shotgun sequencing is primarily a method for studying the functional structure of the communities which aims to examine the entire genetic assemblage and, being amplification-independent, relies on variation and commonality of the collective genomes found in a given environmental sample (Delmont et al. 2012; Prakash & Taylor 2012). Shotgun typing allows for a more comprehensive perspective on the whole microbial community but is limited by its propensity to favour identification of the most dominant members over rarer organisms (Fuhrman 2012). In order to access the rare species found in such a complex matrix as soil, ultra-deep DNA sequencing is required (Howe et al. 2014).

Soil samples obtained during forensic investigations, by their nature, put specific requirements on any metagenomic approach. The main limiting factor of forensic soil samples is their small size and sufficient amount of the sample should remain after analysis for independent re-testing if required. The need for a relatively large amount of initial DNA template for the shotgun sequencing makes this approach less suitable for forensic oriented metagenomic analysis. Multiple displacement amplification using

random hexamers and phage phi29 polymerase (whole genome amplification, WGA) has been reported to successfully amplify minute amounts of DNA in order to produce sufficient quantities recommended for whole-genome shotgun sequencing (Binga et al. 2008).

The ability to identify DNA from the entire genetic composition of soil is desirable for forensic investigation as the DNA from a wide range of organisms may be present: these include the DNA from bacteria, fungi, nematodes, mammals, plant material, and from insect remains. These can be used to generate a rich DNA profile for comparison and meaningful discrimination between soil samples.

Various bioinformatics approaches and tools have been recently developed for description and interpretation of metagenomic sequencing data. For example, taxonomic affiliation and metabolic annotation of sequencing reads derived from the highly diverse soil microbial community can be effectively performed using on-line pipelines with no access to high-performance computers. Online metagenomic annotation services, such as MG-RAST (Meyer et al. 2008), IMG/M (Markowitz et al. 2012) and EMBL-EBI (McWilliam et al. 2013) provide useful means for sequences annotation and gene prediction, assignment of functional categories, description of protein families and genes ontologies, etc .

An outcome of the majority of HTS platforms, with the exception of 454 pyrosequencing technology, is a set of millions of individual reads (sequences) of relatively short length – less than 400 base pairs. These individual reads can be then used to retrieve protein coding regions via processing through on-line metagenomic servers. For example, FragGeneScan algorithm, implemented on the MG-RAST annotation system, allows for annotation of reads starting from those of 75 nt length and longer (Wilke et al. 2013). Assignment of these protein coding regions with reference

databases is the next step of the analysis. Initially genomic databases consisted of reference sequences belonging to easily cultivated microorganisms. However, due to large sequencing projects such as GEBA (Genomic Encyclopaedia of Bacteria and Archaea) (Wu et al. 2009), the number of known sequences included in the databases has been constantly increasing.

Numerous individual reads, some of them are informative by their own, being assembled with other sequencing reads may represent a better picture of the microbial community. The assembly process creates long contiguous sequences, also called as contigs. Contigs of several thousand nucleotides or even complete genomes can be reconstructed depending on the quality of sequencing data and complexity of the initial DNA assemblages. Generation of the assembled reads is performed using different metagenomic assembly programs of which gsAssembler (Margulies et al. 2005), MIRA3 (Biswas et al. n.d.), Meta-IDBA (Peng et al. 2011) and MetaVelvet (Namiki et al. 2012) are the most popular.

The goal of this chapter is to assess the ability SH and WGA sequencing techniques to discriminate visually similar soils of different locality and compare their performance with AP-PCR based sequencing. To achieve this goal the following major tasks were carefully planned and executed:

- Sequencing of metagenomic DNA extracted from similar soil samples from locations A and B, Set 1 described in Chapters 2 & 3, using:
  - shotgun and
  - WGA-based techniques;
  - AP-PCR-based sequencing

- Bioinformatic analysis SH- WGA- and AP-PCR-based datasets (for AP-PCR data taken from Chapter 3) based on (a) full datasets, (b) subsampled datasets and (c) assembled datasets using MG-RAS metagenomic on-line service.
- Comparison of the SH-, WGA- and AP-based taxonomic/metabolic profiles, obtained by mapping against protein M5NR, ribosomal M5RNA and SEED Subsystems databases at all levels of classification available in MG-RAST, using multivariate statistical analysis methods such as:
    - Hierarchical agglomerative clustering (CLUSTER)
    - Non-metric multidimensional scaling (NMDS)
    - Canonical analysis of principal coordinates (CAP).

## 5.2 Materials and Methods

### 5.2.1 DNA specimens.

DNA extracts from soil samples Aw2-Aw4 and Bw2-Bw4 (6 sample in total, for details see Chapter 3 and Chapter 4, Materials and Methods sections) were used for SH and WGA sequencing.

### 5.2.2 Sequencing

For each of the six samples, WGA was conducted with 20 ng DNA using Phi29 DNA polymerase (REPLI-g, Qiagen, Germany). 50 μL amplification reactions were prepared, and each contained 1× reaction buffer, 20 ng of soil DNA extract, 40 units of ϕ29 DNA polymerase (Repli-g kit, Qiagen). Reactions were incubated at either 30°C for 16 h then terminated by heating to 65°C for 5 min. The quality of amplification products was determined by 1% agarose gel electrophoresis and by quantification on a Qubit fluorometer (Life technologies, USA) after purification with a QIAquick PCR Kit (Qiagen, Germany).

Metagenomic library preparation was carried out from 100 ng of DNA extract for SH-based sequencing and 100 ng of WGA amplification products for WGA-based sequencing. Following HTS sequencing was performed at the Australian Genome Research Facility (AGRF, http://www.agrf.org.au/, Adelaide, SA, Australia) using Ion Torrent technology (Ion Torrent PGM Sequencer; Life Technologies, USA) on a separate Ion 318 chip for each of the sequencing approaches (for detailed library preparation and emulsion PCR see Chapters 3 & 4, Materials and Methods sections). Twelve datasets resulted, namely, SH-Aw2 – SH-Aw4, SH-Bw2 – SH-Bw4, WGA-Aw2 – WGA-Aw4 and WGA-Bw2 – WGA-Bw4.

AP-PCR based sequencing datasets AP-Aw2 – AP-Aw4 (soil sample A) and AP-Bw2 – AP-Bw4 (Soil sample B), 6 datasets in total, generated in Chapter 4 were also employed for bioinformatics processing, reference database annotation and statistical comparative analysis.

### 5.2.3 Processing of sequencing data

All datasets from SH-, WGA- and AP-PCR based sequencing were processed using the same metagenomic pipeline under the same conditions. Raw sequence datasets were uploaded to the Metagenome Rapid Annotation using Subsystem Technology (MG-RAST) server (http://metagenomics.nmpdr.org/) (Meyer et al., 2008) and filtered from low-quality reads prior to annotation. Metagenomic datasets were annotated to protein genes against the M5NR database and SEED Subsystems database resulting in protein-derived taxonomic and metabolic profiles, respectively. In addition taxonomic profiles were generated by comparison of the metagenomic datasets with the M5RNA ribosomal database also available in MG-RAST. The MG-RAST default annotation parameters, such as maximum E-value < $1x10^{-5}$, minimum length of alignment of 15 amino acids for protein database annotation and 15 bp for rRNA database annotation along with minimum sequence identity of 60%, were used to identify the best database matches. Metagenomic profiles were generated at all available MG-RAST taxonomic (Phylum to species) and metabolic (level 1 to functions) levels of hierarchy. To adjust the differences in sequencing effort across samples, two common procedure of standardization were taken:

1. In the first approach metagenomic profiles were generated using full datasets of the high-quality reads obtained for each sample. For the metagenomic comparison of profiles the relative abundance scores for each taxon and metabolic feature were

determined by the percentages of respective reads over the total number of assigned reads. In the text the relative abundance scores found both for the taxonomic and metabolic features are represented as an average ± SD (standard deviation) across all datasets (if not mentioned otherwise).

2. A second approach was based on comparison of metagenomic profiles generated from randomly subsampled datasets of 49 000 annotated reads per sample.

*Annotation of the assembled metagenomic sequence datasets*

Obtained datasets were assembled with the MetaVelvet assembler (v.1.2.01) with the following parameters: k-mer length (K) 59 and automatic identification of the expected coverage value (exp_cov auto).

Assembled contigs and their corresponding median base pair coverage were uploaded into the MG-RAST annotation pipeline, using the same settings as for the annotation of the raw sequence datasets, such as a maximum E-value of $< 1x10^{-5}$, a minimum identity of 60%, and a minimum alignment length of 15 amino acids for the comparison with protein M5NR, SEED Subsystems and 15 bp for M5RNA reference databases. Individual reads and assembled metagenomic datasets with their MG-RAST sample IDs are listed in the Table 5.1.

**Table 5.1. Summary of soil metagenomic datasets.** Datasets are publically available on the MG-RAST server (http://metagenomics.anl.gov/). SH = shotgun, AP = arbitrary primed PCR, WGA = whole genome amplification.

| Group | Full sequencing datasets | Randomly subsampled datasets | Assembled datasets |
|---|---|---|---|
| **SH_A** | 4533948.3 | 4553173.3 | 4600135.3 |
| | 4533949.3 | 4553184.3 | 4600136.3 |
| | 4533950.3 | 4553185.3 | 4600137.3 |
| **SH_B** | 4533951.3 | 4553186.3 | 4600138.3 |
| | 4533952.3 | 4553187.3 | 4600139.3 |
| | 4533953.3 | 4553188.3 | 4600140.3 |
| **AP_A** | 4549132.3 | 4553174.3 | 4600103.3 |
| | 4549136.3 | 4553178.3 | 4600107.3 |
| | 4549137.3 | 4553179.3 | 4600108.3 |
| **AP_B** | 4549141.3 | 4553180.3 | 4600113.3 |
| | 4549142.3 | 4553181.3 | 4600114.3 |
| | 4549144.3 | 4553183.3 | 4600116.3 |
| **WGA_A** | 4543715.3 | 4553189.3 | 4600141.3 |
| | 4543716.3 | 4553190.3 | 4600142.3 |
| | 4543717.3 | 4553191.3 | 4600143.3 |
| **WGA_B** | 4543718.3 | 4553192.3 | 4600144.3 |
| | 4543719.3 | 4553193.3 | 4600145.3 |
| | 4543720.3 | 4553194.3 | 4600146.3 |

### 5.2.4 Statistical analysis of data.

Statistical approaches and tools used for the analysis of the obtained metagenomic data were as per described in Chapters 2 – 4 such as CLUSTER, NMDS, SIMPROF and CAP. Additionally, RELATE and Rarefaction analyses were performed.

The species richness was estimated by rarefaction analysis preformed in MG-RAST. The analysis was performed for total taxa identified with the M5NR protein database in randomly subsampled metagenomic datasets (including Bacteria, Archaea, Eukaryota, Viruses, unclassified and other sequences).

The program RELATE in the PRIMER 6 package was used to calculate the Spearman rank correlation between Bray-Curtis similarity matrices generated for full and subsampled datasets (Clarke & Warwick 2001).

## 5.3 Results

### 5.3.1 Notation and general characteristics of sequencing datasets.

For each soil DNA sample three datasets were generated from the same DNA template using three sequencing approaches, namely shotgun (SH), whole genome amplification (WGA) and arbitrary primed PCR (AP-PCR). SH sequencing resulted in an average of 672,542 (531,108 – 806,843) sequence reads with an average sequence length of 198 ± 73 bases for a total of > 133 Mbp of sequence. WGA dataset consisted of an average of 911,554 (506,028 – 2,012,359) sequences with an average of 198 ± 75 bases in length for a total of > 178 Mbp. The AP-based approach gave an average of 468,187 (74,370 – 1,047,266) reads with an average 143 ± 69 bases in length for a total of > 70.7Mbp (Table 5.2), as was described in Chapter 4.

**Table 5.2. General characteristics of sequence data obtained.**

| Sequencing approach | Average number of reads (range) | Number of Mbp | Average read length, bp ± SD | Failed QF, % | Average number of contigs | Number of Mbp | Average read length, bp ± SD (max length) | Failed QF, % |
|---|---|---|---|---|---|---|---|---|
| SH | 672,542 (531,108-806,483) | 133.6 | 198 ± 73 | 21 | 16,665 (9,871 – 20,893) | 3.7 | 223 ± 40 (773 ±137) | 7 |
| AP | 468,187 (74,370-1,074,266) | 70.7 | 142 ± 69 | 24 | 12,148 (712 – 27,497) | 2.1 | 202 ± 114 (1240 ± 116) | 18 |
| WGA | 911,553 (506,028-2,012,359) | 178.5 | 198 ± 75 | 20 | 40,794 (9,946 - 89,507) | 9.3 | 233 ± 92 (1882 ± 724) | 13 |

Statistical data represented as mean ± Standard Deviation (SD). QF = Quality Filtering.

**Table 5.3. Characteristics of MG-RAST annotation of full sequence data.**

| Sequencing Approach | Number of reads passed QF (range) | Number of predicted protein features (range) | Number of predicted rRNA features (range) | Number of protein features assigned to M5NR database (%) | Number of protein features assigned to SEED Subsystems database (%) | Number of ribosomal features assigned to M5RNA (%) |
|---|---|---|---|---|---|---|
| SH | 536,960 (385,996 – 661,193) | 440,507 (310,581 – 551,558) | 82,151 (62,899 – 96,886) | 35 | 43 | 1.3 |
| AP | 352,265 (56,461 – 774,443) | 287,840 (49,902 – 617,609) | 44,896 (7,295 – 104,247) | 26 | 30 | 0.0 |
| WGA | 716,902 (407,932 – 1,521,697) | 439,129 (343,052 – 689,805) | 96,117 (61,694 – 187,539) | 26 | 31 | 0.8 |

Percentage of sequences matching to the M5NR, M5RNA and SEED Subsystems databases was determined with an E-value cut-off of $E<1x10^{-5}$.

**Table 5.4. Characteristics of MG-RAST annotation of assembled sequence data.**

| Sequencing Approach | Number of contigs passed QF (range) | Number of predicted protein features (range) | Number of predicted rRNA features (range) | Number of protein features assigned to M5NR database, % | Number of protein features assigned to SEED Subsystems database, % | Number of ribosomal features assigned to M5RNA, % |
|---|---|---|---|---|---|---|
| SH | 15,459 (9,124 – 19,510) | 14,932 (8,845 – 19,046) | 322 (235 – 421) | 43 | 54 | 68.5 |
| AP | 10,112 (547 – 23,020) | 8,463 (540 – 19,183) | 60 (0 – 126) | 26 | 36 | 0.0 |
| WGA | 34,988 (9,147 – 76,561) | 31,170 (8,930 – 65,022) | 611 (163 – 1 886) | 26 | 27 | 40.1 |

From the SH dataset on average 16,665 (9,871 – 20,893) contigs with maximum length of 773 ± 137 bp were generated by MetaVelvet assembling software. The WGA-based and AP-based datasets resulted in an average of 40,794 (9,946 – 89,507) and 12,148 (712 – 27,497) contigs with maximum length of 1,882 ± 724 and 1,240 ± 116 bp, respectively (Table 5.2).

Datasets were then uploaded to the online MG-RAST server (Meyer et al. 2008) for the quality filtering (QF) and annotation with different reference databases. Approximately 22% of low quality reads were eliminated from each of initial datasets at the filtering step and 13% from assembled datasets (Table 5.2). Quality filtering procedure insures that no reads containing more than five bases with Phred score less than 15 (MG-RAST default settings) are included in the consecutive analysis. Phred score 15 corresponds to the base calling accuracy of 97%. Annotation of quality filtered initial and assembled datasets was performed using protein and ribosomal reference databases. Only 25–35% of the protein features across all shotgun, WGA-based and AP-based sequence datasets which contained predicted protein coding regions (49,902 – 689,805 reads per sample), were taxonomically assigned using M5NR protein database. While 30 – 40% of protein features assigned to the SEED Subsystems database were used for the generation of metabolic profiles (Table 5.3). Comparison of assembled datasets with M5NR database resulted in 25-40% of predicted protein features (540 –

65,022) to be assigned and 27-54% were assigned to SEED Subsystems database (Table 5.4).

Each of the initial metagenomic datasets, according to the MG-RAST statistics, contained from 7,295 to 187,539 features predicted to be rRNA. However, the subsequent annotation revealed no identified rRNA features in the AP-based dataset and only 1% of the predicted rRNA features from the SH- and WGA-based datasets matched the M5RNA database (Table 5.3). In the assembled datasets, a substantially lower number of predicted rRNA features were found (0 – 1,886) (Table 5.4). None of the sequences from AP-based dataset matched any sequence on the M5RNA database, however 40% and 68% of predicted rRNA features from WGA-based and SH-based were identified within the M5RNA database, respectively.

### 5.3.2 Taxonomic profiling of metagenomes

Analyses of metagenomic data within MG-RAST occurs both for protein coding genes and ribosomal (rRNA) genes. The analysis of taxonomy can therefore be performed in two ways.

Taxonomic classification of protein gene fragments showed that 85 ± 4% of annotated reads and assembled contigs were assigned to Bacteria. On average, 3.9 ± 2.5% of reads and 4.3 ± 2.7% of assembled contigs matched to Eukaryota. To Archaea domain, 0.6 ± 0.4% and 0.6 ± 0.3% of reads and assembled contigs were assigned, respectively. The remaining 10.8 ± 1.7% of reads and 8.7 ± 1.4% of the assembled contigs were not assigned. Bacterial taxa *Proteobacteria*, *Actinobacteria* and *Bacteroidetes* dominated in all metagenomic datasets, including assembled contigs, representing close to 70% of protein annotated reads. Additional phyla including

*Chloroflexi, Planctomycetes, Acidobacteria Firmicutes, Cyanobacteria, Verrucomicrobia* represented less than 5% of reads and contigs. *Ascomycota* was found to be the dominant microorganism with 3.0 ± 2.6% of reads and 3.6 ± 2.4% assembled contigs Other eukaryotic taxa such as *Streptophyta, Chordata, Basidiomycota* and *Arthropoda* collectively contributed to the remaining 1% of the annotated reads and contigs (Table 5.5 and Table 5.7).

Taxonomic classification of the rRNA gene fragments identified only in SH- and WGA-based datasets showed that 78 ± 8% of reads and 75.5 ± 4.6% of the assembled contigs were assigned to bacterial taxa. Eukaryotic taxa were found in 14.5 ± 6.5% of reads and 6.9 ± 3.2% of contigs (data represented as an average relative abundance of taxa between the samples of SH- and WGA-based datasets). The most abundant bacterial and eukaryotic phyla found were the same as per protein-derived taxonomic classification (described above) namely: *Actinobacteria, Proteobacteria, Bacteroidetes, Ascomycota* and *Streptophyta.* The remaining 15.8 ± 5.4% of reads and 15.9 ± 4.9% of the assembled contigs were not assigned (Table 5.6 and Table 5.8).

**Table 5.5. Protein-derived taxonomic composition of the soil microbial communities built on the initial sequencing datasets.** Relative abundances of major taxa (phylum level) derived from taxonomic assignment of protein gene fragments matched to M5NR database.

| | M5NR | WGA | | | | | | SH | | | | | | AP | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **domain** | **phylum** | **Aw2** | **Aw3** | **Aw4** | **Bw2** | **Bw3** | **Bw4** | **Aw2** | **Aw3** | **Aw4** | **Bw2** | **Bw3** | **Bw4** | **Aw2** | **Aw3** | **Aw4** | **Bw2** | **Bw3** | **Bw4** |
| Archaea | Thaumarchaeota | 0.12 | 0.27 | 0.62 | 0.29 | 0.02 | 0.05 | 0.22 | 0.09 | 0.43 | 0.04 | 0.19 | 0.13 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 |
| | Crenarchaeota | 0.06 | 0.13 | 0.22 | 0.13 | 0.03 | 0.06 | 0.11 | 0.06 | 0.17 | 0.05 | 0.09 | 0.08 | 0.03 | 0.05 | 0.12 | 0.01 | 0.01 | 0.03 |
| | Euryarchaeota | 0.36 | 0.59 | 0.78 | 0.48 | 0.30 | 0.28 | 0.42 | 0.29 | 0.51 | 0.30 | 0.40 | 0.36 | 0.77 | 0.26 | 0.45 | 0.18 | 0.20 | 0.12 |
| | Total | 0.54 | 1.00 | 1.62 | 0.90 | 0.35 | 0.39 | 0.75 | 0.44 | 1.12 | 0.40 | 0.69 | 0.57 | 0.80 | 0.31 | 0.57 | 0.19 | 0.22 | 0.15 |
| Bacteria | Verrucomicrobia | 1.49 | 2.19 | 1.83 | 3.02 | 2.50 | 1.56 | 1.64 | 1.21 | 1.38 | 2.43 | 2.03 | 1.47 | 1.27 | 2.07 | 1.24 | 1.86 | 2.55 | 1.90 |
| | Bacteroidetes | 24.65 | 21.38 | 18.46 | 20.87 | 8.56 | 5.22 | 8.63 | 11.75 | 6.94 | 13.59 | 8.80 | 8.28 | 9.91 | 9.57 | 4.43 | 5.04 | 6.67 | 6.37 |
| | Actinobacteria | 10.67 | 11.92 | 14.46 | 8.93 | 28.99 | 43.12 | 32.43 | 27.16 | 35.10 | 20.20 | 26.89 | 35.89 | 12.16 | 12.48 | 23.71 | 9.92 | 6.93 | 11.94 |
| | Chlorobi | 0.43 | 0.60 | 0.70 | 0.50 | 0.25 | 0.18 | 0.27 | 0.25 | 0.26 | 0.31 | 0.34 | 0.26 | 0.31 | 0.14 | 0.17 | 1.74 | 1.20 | 1.27 |
| | Nitrospirae | 0.10 | 0.23 | 0.22 | 0.35 | 0.11 | 0.09 | 0.14 | 0.09 | 0.13 | 0.12 | 0.23 | 0.13 | 0.08 | 0.07 | 0.16 | 0.14 | 0.10 | 0.14 |
| | Deinococcus-Thermus | 0.37 | 0.46 | 0.55 | 0.40 | 0.36 | 0.37 | 0.45 | 0.32 | 0.54 | 0.35 | 0.39 | 0.38 | 0.54 | 0.31 | 0.50 | 0.27 | 0.48 | 0.45 |
| | Gemmatimonadetes | 0.10 | 0.15 | 0.17 | 0.22 | 0.23 | 0.23 | 0.24 | 0.14 | 0.30 | 0.24 | 0.36 | 0.23 | 0.57 | 0.43 | 0.28 | 3.73 | 2.36 | 3.97 |
| | Acidobacteria | 1.22 | 1.57 | 1.60 | 2.22 | 1.60 | 1.14 | 1.41 | 1.00 | 1.31 | 1.79 | 1.73 | 1.25 | 5.89 | 4.29 | 4.08 | 6.15 | 4.62 | 4.89 |
| | Spirochaetes | 0.16 | 0.19 | 0.18 | 0.21 | 0.13 | 0.11 | 0.15 | 0.12 | 0.14 | 0.16 | 0.19 | 0.14 | 0.03 | 0.07 | 0.06 | 0.01 | 0.07 | 0.03 |
| | Firmicutes | 2.43 | 3.26 | 3.84 | 2.81 | 2.17 | 2.03 | 2.72 | 2.13 | 2.86 | 2.30 | 2.73 | 2.37 | 2.96 | 1.38 | 2.12 | 0.79 | 0.68 | 1.06 |
| | Chloroflexi | 1.19 | 1.86 | 2.60 | 1.12 | 0.86 | 0.91 | 1.47 | 0.91 | 1.88 | 0.87 | 1.03 | 1.01 | 1.56 | 5.15 | 3.43 | 6.26 | 6.91 | 5.34 |
| | Planctomycetes | 1.22 | 2.65 | 2.27 | 3.42 | 2.92 | 2.42 | 2.86 | 1.42 | 2.56 | 3.32 | 3.50 | 2.76 | 0.95 | 5.65 | 4.36 | 1.81 | 1.84 | 2.06 |
| | Proteobacteria | 31.00 | 33.46 | 33.79 | 33.21 | 35.77 | 29.70 | 33.66 | 38.20 | 33.24 | 37.11 | 36.87 | 31.94 | 39.16 | 41.50 | 38.47 | 47.02 | 47.66 | 45.15 |
| | Chlamydiae | 0.06 | 0.08 | 0.10 | 0.10 | 0.06 | 0.05 | 0.06 | 0.05 | 0.06 | 0.10 | 0.07 | 0.07 | 0.02 | 0.03 | 0.03 | 0.02 | 0.00 | 0.02 |
| | Cyanobacteria | 1.35 | 2.14 | 2.48 | 1.76 | 1.08 | 0.91 | 1.29 | 0.97 | 1.46 | 1.15 | 1.20 | 1.03 | 1.22 | 1.42 | 1.36 | 1.03 | 1.34 | 1.04 |
| | Total | 76.44 | 82.15 | 83.25 | 79.15 | 85.59 | 88.04 | 87.42 | 85.71 | 88.15 | 84.02 | 86.34 | 87.19 | 76.63 | 84.56 | 84.39 | 85.78 | 83.40 | 85.64 |
| Eukaryota | Ascomycota | 10.56 | 3.74 | 2.32 | 6.54 | 3.53 | 1.86 | 1.01 | 2.97 | 0.58 | 3.30 | 1.99 | 1.85 | 5.77 | 1.86 | 0.94 | 0.57 | 2.29 | 1.59 |
| | Streptophyta | 0.20 | 0.25 | 0.21 | 0.20 | 0.16 | 0.13 | 0.16 | 0.16 | 0.15 | 0.16 | 0.17 | 0.15 | 0.29 | 1.41 | 0.43 | 1.47 | 2.16 | 1.80 |
| | Chordata | 0.30 | 0.45 | 0.31 | 0.29 | 0.19 | 0.28 | 0.23 | 0.20 | 0.16 | 0.18 | 0.20 | 0.30 | 0.08 | 0.03 | 0.12 | 0.01 | 0.04 | 0.02 |
| | Basidiomycota | 0.30 | 0.17 | 0.13 | 0.22 | 0.10 | 0.07 | 0.06 | 0.08 | 0.05 | 0.10 | 0.09 | 0.08 | 0.19 | 0.05 | 0.13 | 0.01 | 0.04 | 0.06 |
| | Arthropoda | 0.18 | 0.45 | 0.13 | 0.12 | 0.09 | 0.16 | 0.25 | 0.16 | 0.06 | 0.09 | 0.08 | 0.24 | 0.05 | 0.03 | 0.06 | 0.00 | 0.03 | 0.06 |
| | Total | 11.54 | 5.06 | 3.10 | 7.36 | 4.07 | 2.50 | 1.72 | 3.56 | 1.00 | 3.84 | 2.52 | 2.62 | 6.38 | 3.38 | 1.67 | 2.06 | 4.56 | 3.53 |
| | Other | 0.65 | 0.90 | 0.86 | 0.86 | 0.60 | 0.60 | 0.64 | 0.51 | 0.64 | 0.65 | 0.69 | 0.63 | 0.20 | 0.28 | 0.68 | 0.12 | 0.22 | 0.19 |
| | Unassigned | 10.82 | 10.89 | 11.17 | 11.72 | 9.39 | 8.46 | 9.47 | 9.79 | 9.09 | 11.09 | 9.76 | 8.99 | 15.99 | 11.47 | 12.68 | 11.85 | 11.59 | 10.48 |

**Here and thereafter red shaded cells show the highest value within corresponding group**

**Table 5.6. rRNA-based taxonomic composition of the soil microbial communities built on the initial sequencing datasets.** Relative abundances of major taxa (phylum level) derived from taxonomic assignment of ribosomal gene fragments matched to M5RNA database.

| domain | M5RNA phylum | WGA Aw2 | WGA Aw3 | WGA Aw4 | WGA Bw2 | WGA Bw3 | WGA Bw4 | SH Aw2 | SH Aw3 | SH Aw4 | SH Bw2 | SH Bw3 | SH Bw4 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Archaea | Thaumarchaeota | 0.11 | 0.25 | 0.00 | 0.06 | 0.00 | 0.00 | 0.06 | 0.04 | 0.32 | 0.06 | 0.16 | 0.16 |
| | Crenarchaeota | 0.00 | 0.13 | 0.31 | 0.31 | 0.00 | 0.11 | 0.00 | 0.00 | 0.13 | 0.00 | 0.00 | 0.00 |
| | Euryarchaeota | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.08 | 0.00 |
| | Total | 0.11 | 0.38 | 0.31 | 0.37 | 0.00 | 0.11 | 0.06 | 0.04 | 0.45 | 0.06 | 0.23 | 0.16 |
| Bacteria | Verrucomicrobia | 0.46 | 0.38 | 0.31 | 0.62 | 0.85 | 0.65 | 0.24 | 0.46 | 0.39 | 1.51 | 0.23 | 0.78 |
| | Bacteroidetes | 10.07 | 7.50 | 5.52 | 2.11 | 4.26 | 1.08 | 1.83 | 3.50 | 0.97 | 5.36 | 1.72 | 1.49 |
| | Actinobacteria | 21.68 | 29.99 | 39.26 | 37.38 | 47.87 | 61.64 | 55.22 | 49.68 | 61.02 | 42.67 | 63.81 | 53.65 |
| | Chlorobi | 0.00 | 0.00 | 0.00 | 0.12 | 0.00 | 0.11 | 0.06 | 0.00 | 0.00 | 0.00 | 0.00 | 0.08 |
| | Nitrospirae | 0.07 | 0.00 | 0.00 | 0.19 | 0.00 | 0.00 | 0.06 | 0.00 | 0.13 | 0.06 | 0.08 | 0.08 |
| | Deinococcus-Thermus | 0.11 | 0.00 | 0.00 | 0.06 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Gemmatimonadetes | 0.00 | 0.00 | 0.00 | 0.06 | 0.00 | 0.00 | 0.12 | 0.15 | 0.13 | 0.00 | 0.00 | 0.00 |
| | Acidobacteria | 0.61 | 0.38 | 0.00 | 0.99 | 0.34 | 0.00 | 0.06 | 0.23 | 0.00 | 0.47 | 0.08 | 0.16 |
| | Spirochaetes | 0.11 | 0.38 | 0.15 | 0.19 | 0.00 | 0.22 | 0.12 | 0.04 | 0.32 | 0.00 | 0.00 | 0.16 |
| | Firmicutes | 1.07 | 1.02 | 2.45 | 1.05 | 1.19 | 0.65 | 1.53 | 0.95 | 1.55 | 0.58 | 0.47 | 1.49 |
| | Chloroflexi | 0.07 | 0.38 | 0.46 | 0.25 | 0.17 | 0.32 | 0.37 | 0.04 | 0.65 | 0.17 | 0.23 | 0.31 |
| | Planctomycetes | 0.18 | 0.64 | 0.46 | 0.62 | 1.53 | 0.54 | 1.34 | 0.65 | 0.90 | 0.81 | 0.55 | 2.12 |
| | Proteobacteria | 32.04 | 27.45 | 25.61 | 18.66 | 13.97 | 11.85 | 15.39 | 16.48 | 13.32 | 14.67 | 9.75 | 15.22 |
| | Chlamydiae | 0.07 | 0.00 | 0.00 | 0.25 | 0.00 | 0.00 | 0.12 | 0.08 | 0.06 | 0.17 | 0.00 | 0.00 |
| | Cyanobacteria | 0.14 | 0.00 | 0.31 | 0.00 | 0.34 | 0.11 | 0.18 | 0.11 | 0.06 | 0.12 | 0.00 | 0.00 |
| | Total | 66.68 | 68.11 | 74.54 | 62.55 | 70.53 | 77.16 | 76.66 | 72.36 | 79.51 | 66.59 | 76.91 | 75.53 |
| Eukaryota | Ascomycota | 5.96 | 3.81 | 1.84 | 5.46 | 9.03 | 3.56 | 3.79 | 8.76 | 3.36 | 11.82 | 4.84 | 5.33 |
| | Streptophyta | 1.71 | 2.54 | 0.92 | 2.05 | 2.73 | 1.94 | 0.98 | 1.37 | 1.16 | 1.86 | 0.94 | 0.63 |
| | Chordata | 0.29 | 0.51 | 0.31 | 0.74 | 3.75 | 3.99 | 0.61 | 0.34 | 1.10 | 1.05 | 0.31 | 0.55 |
| | Basidiomycota | 0.64 | 0.13 | 0.77 | 0.68 | 0.51 | 0.22 | 0.37 | 0.69 | 0.26 | 0.64 | 0.47 | 0.71 |
| | Arthropoda | 0.32 | 1.14 | 0.46 | 0.19 | 0.51 | 2.80 | 1.71 | 0.88 | 0.97 | 0.76 | 2.89 | 0.39 |
| | Total | 8.93 | 8.13 | 4.29 | 9.11 | 16.52 | 12.50 | 7.45 | 12.03 | 6.85 | 16.12 | 9.44 | 7.61 |
| | Other | 1.04 | 2.29 | 1.38 | 2.36 | 0.85 | 1.94 | 3.60 | 0.65 | 1.55 | 1.86 | 2.57 | 1.49 |
| | Unclassified | 23.25 | 21.09 | 19.33 | 25.60 | 12.10 | 8.30 | 12.22 | 14.92 | 11.64 | 15.37 | 10.84 | 15.14 |

Annotation revealed no identified rRNA features in the AP-based datasets

**Table 5.7. Protein-derived taxonomic composition of the soil microbial communities built on the assembled contigs.** Relative abundances of major taxa (phylum level) derived from taxonomic assignment of protein gene fragments matched to M5NR database.

| M5NR | | WGA | | | | | | SH | | | | | | AP | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **domain** | **phylum** | **Aw2** | **Aw3** | **Aw4** | **Bw2** | **Bw3** | **Bw4** | **Aw2** | **Aw3** | **Aw4** | **Bw2** | **Bw3** | **Bw4** | **Aw2** | **Aw3** | **Aw4** | **Bw2** | **Bw3** | **Bw4** |
| Archaea | Thaumarchaeota | 0.05 | 0.20 | 0.38 | 0.32 | 0.03 | 0.05 | 0.27 | 0.05 | 0.62 | 0.06 | 0.30 | 0.21 | 0.03 | 0.03 | 0.03 | 0.00 | 0.00 | 0.00 |
| | Crenarchaeota | 0.05 | 0.05 | 0.17 | 0.16 | 0.01 | 0.04 | 0.20 | 0.05 | 0.16 | 0.03 | 0.18 | 0.06 | 0.07 | 0.08 | 0.06 | 0.00 | 0.45 | 0.00 |
| | Euryarchaeota | 0.66 | 0.97 | 1.35 | 0.47 | 0.23 | 0.25 | 0.53 | 0.27 | 0.54 | 0.25 | 0.52 | 0.43 | 0.65 | 0.68 | 0.90 | 0.71 | 0.91 | 0.29 |
| | Total | 0.75 | 1.21 | 1.90 | 0.95 | 0.27 | 0.33 | 1.01 | 0.38 | 1.32 | 0.34 | 1.00 | 0.70 | 0.75 | 0.78 | 1.00 | 0.71 | 1.36 | 0.29 |
| Bacteria | Verrucomicrobia | 1.07 | 1.16 | 1.16 | 2.77 | 1.87 | 1.13 | 1.43 | 0.99 | 1.28 | 2.17 | 1.74 | 1.29 | 2.47 | 3.87 | 1.81 | 3.66 | 4.99 | 3.73 |
| | Bacteroidetes | 16.97 | 10.79 | 7.67 | 17.60 | 8.19 | 3.85 | 8.75 | 11.73 | 6.56 | 14.98 | 9.99 | 8.65 | 5.81 | 5.35 | 2.53 | 5.18 | 3.85 | 4.22 |
| | Actinobacteria | 12.08 | 12.73 | 15.98 | 7.15 | 33.97 | 51.14 | 33.77 | 28.26 | 37.39 | 19.81 | 29.34 | 38.91 | 14.77 | 19.18 | 25.73 | 8.74 | 10.20 | 12.57 |
| | Chlorobi | 0.60 | 0.88 | 0.93 | 0.59 | 0.21 | 0.19 | 0.32 | 0.23 | 0.25 | 0.34 | 0.45 | 0.13 | 0.30 | 0.27 | 0.15 | 0.71 | 0.00 | 0.29 |
| | Nitrospirae | 0.16 | 0.31 | 0.21 | 0.35 | 0.10 | 0.10 | 0.16 | 0.07 | 0.11 | 0.08 | 0.22 | 0.15 | 0.24 | 0.27 | 0.25 | 0.30 | 0.45 | 0.39 |
| | Deinococcus-Thermus | 0.39 | 0.43 | 0.57 | 0.46 | 0.26 | 0.36 | 0.43 | 0.28 | 0.57 | 0.30 | 0.49 | 0.52 | 0.64 | 0.63 | 0.69 | 1.83 | 0.68 | 0.49 |
| | Gemmatimonadetes | 0.10 | 0.17 | 0.14 | 0.15 | 0.14 | 0.17 | 0.24 | 0.19 | 0.23 | 0.29 | 0.32 | 0.27 | 0.31 | 0.47 | 0.46 | 1.32 | 0.23 | 0.39 |
| | Acidobacteria | 1.13 | 1.25 | 1.34 | 2.12 | 1.38 | 1.07 | 1.35 | 0.98 | 1.09 | 1.68 | 1.47 | 0.94 | 2.20 | 2.59 | 2.14 | 3.66 | 3.17 | 3.24 |
| | Spirochaetes | 0.12 | 0.11 | 0.11 | 0.18 | 0.09 | 0.07 | 0.18 | 0.13 | 0.14 | 0.15 | 0.21 | 0.17 | 0.07 | 0.28 | 0.20 | 0.10 | 0.00 | 0.29 |
| | Firmicutes | 3.23 | 4.17 | 5.17 | 3.07 | 1.80 | 1.62 | 3.19 | 1.97 | 3.41 | 2.40 | 3.03 | 2.06 | 2.69 | 2.86 | 2.77 | 3.25 | 1.36 | 2.26 |
| | Chloroflexi | 2.22 | 3.62 | 3.65 | 1.21 | 0.67 | 0.77 | 1.39 | 1.04 | 2.07 | 1.01 | 0.99 | 1.29 | 2.37 | 3.26 | 3.68 | 2.85 | 2.27 | 2.75 |
| | Planctomycetes | 1.04 | 2.00 | 1.53 | 3.11 | 2.26 | 1.82 | 2.69 | 1.15 | 2.35 | 3.03 | 3.41 | 2.74 | 2.53 | 5.10 | 5.39 | 5.69 | 6.12 | 5.30 |
| | Proteobacteria | 38.05 | 41.62 | 42.04 | 40.22 | 35.79 | 26.85 | 32.94 | 38.41 | 31.80 | 37.39 | 34.27 | 29.96 | 41.41 | 37.48 | 37.30 | 44.21 | 44.44 | 42.04 |
| | Chlamydiae | 0.05 | 0.09 | 0.11 | 0.14 | 0.03 | 0.03 | 0.06 | 0.07 | 0.07 | 0.07 | 0.09 | 0.09 | 0.07 | 0.13 | 0.06 | 0.00 | 0.00 | 0.10 |
| | Cyanobacteria | 2.53 | 3.75 | 4.01 | 2.54 | 0.94 | 0.76 | 1.37 | 1.24 | 1.67 | 1.23 | 1.33 | 1.18 | 1.59 | 2.00 | 1.98 | 1.22 | 0.68 | 1.28 |
| | Total | 79.73 | 83.08 | 84.62 | 81.66 | 87.69 | 89.92 | 88.26 | 86.74 | 89.00 | 84.92 | 87.33 | 88.35 | 77.45 | 83.72 | 85.13 | 82.72 | 78.46 | 79.37 |
| Eukaryota | Ascomycota | 8.17 | 2.22 | 1.65 | 4.80 | 3.09 | 1.42 | 0.99 | 3.27 | 0.65 | 3.79 | 2.01 | 2.05 | 9.09 | 3.68 | 1.26 | 3.05 | 6.12 | 7.56 |
| | Streptophyta | 0.17 | 0.22 | 0.29 | 0.16 | 0.15 | 0.12 | 0.17 | 0.20 | 0.13 | 0.27 | 0.18 | 0.08 | 0.26 | 0.20 | 0.31 | 0.41 | 0.91 | 0.20 |
| | Chordata | 0.26 | 0.32 | 0.20 | 0.16 | 0.24 | 0.54 | 0.24 | 0.21 | 0.19 | 0.27 | 0.18 | 0.25 | 0.15 | 0.16 | 0.17 | 0.00 | 0.00 | 0.39 |
| | Basidiomycota | 0.23 | 0.04 | 0.10 | 0.19 | 0.05 | 0.05 | 0.04 | 0.13 | 0.04 | 0.10 | 0.08 | 0.09 | 0.63 | 0.16 | 0.05 | 0.00 | 0.68 | 0.10 |
| | Arthropoda | 0.16 | 0.37 | 0.09 | 0.09 | 0.11 | 0.16 | 0.31 | 0.10 | 0.03 | 0.12 | 0.12 | 0.24 | 0.06 | 0.13 | 0.03 | 0.00 | 0.00 | 0.20 |
| | Total | 8.99 | 3.16 | 2.33 | 5.42 | 3.64 | 2.30 | 1.74 | 3.90 | 1.04 | 4.55 | 2.56 | 2.71 | 10.19 | 4.34 | 1.81 | 3.46 | 7.71 | 8.45 |
| | Other | 1.81 | 2.59 | 2.16 | 1.78 | 1.17 | 1.16 | 1.64 | 1.18 | 1.36 | 1.31 | 1.41 | 1.35 | 1.76 | 1.84 | 2.24 | 2.64 | 1.36 | 2.55 |
| | Unassigned | 8.72 | 9.94 | 8.90 | 10.21 | 7.23 | 6.29 | 7.34 | 7.78 | 7.25 | 8.88 | 7.69 | 6.89 | 9.84 | 9.33 | 9.83 | 10.47 | 11.11 | 9.33 |

**Table 5.8. rRNA-based taxonomic composition of the soil microbial communities built on the assembled contigs.** Relative abundances of major taxa (phylum level) derived from taxonomic assignment of ribosomal gene fragments matched to M5RNA database.

| | M5RNA | WGA | | | | | | SH | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **domain** | **phylum** | **Aw2** | **Aw3** | **Aw4** | **Bw2** | **Bw3** | **Bw4** | **Aw2** | **Aw3** | **Aw4** | **Bw2** | **Bw3** | **Bw4** |
| Archaea | Thaumarchaeota | 0.00 | 0.00 | 0.59 | 0.31 | 0.00 | 0.00 | 0.00 | 0.00 | 0.36 | 0.34 | 0.00 | 0.00 |
| | Crenarchaeota | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.03 |
| | Euryarchaeota | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.36 | 0.00 | 0.00 | 0.00 |
| | Total | 0.00 | 0.00 | 0.59 | 0.31 | 0.00 | 0.00 | 0.00 | 0.00 | 0.72 | 0.34 | 0.00 | 1.03 |
| Bacteria | Verrucomicrobia | 0.16 | 0.00 | 0.59 | 0.31 | 0.00 | 0.00 | 0.00 | 0.26 | 0.00 | 0.68 | 1.60 | 1.03 |
| | Bacteroidetes | 9.00 | 14.29 | 4.71 | 3.75 | 1.09 | 0.74 | 1.31 | 5.63 | 2.17 | 7.14 | 0.53 | 2.58 |
| | Actinobacteria | 24.80 | 26.71 | 40.59 | 41.88 | 56.52 | 73.53 | 61.57 | 46.80 | 64.26 | 46.26 | 61.50 | 58.76 |
| | Chlorobi | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.26 | 0.00 | 0.00 | 0.53 | 0.00 |
| | Nitrospirae | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.52 |
| | Deinococcus-Thermus | 0.16 | 0.62 | 0.00 | 0.31 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Gemmatimonadetes | 0.00 | 0.00 | 0.00 | 0.31 | 0.00 | 0.00 | 0.00 | 0.26 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Acidobacteria | 0.63 | 0.00 | 0.00 | 0.63 | 0.00 | 0.00 | 0.00 | 0.26 | 0.72 | 1.36 | 0.00 | 0.00 |
| | Spirochaetes | 0.16 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.53 | 0.00 |
| | Firmicutes | 2.21 | 1.86 | 1.76 | 0.63 | 1.09 | 0.00 | 0.44 | 1.79 | 1.81 | 1.02 | 1.07 | 0.52 |
| | Chloroflexi | 0.00 | 0.00 | 2.35 | 0.63 | 1.09 | 0.00 | 0.00 | 0.00 | 0.72 | 0.00 | 0.00 | 0.52 |
| | Planctomycetes | 0.32 | 1.24 | 0.00 | 0.63 | 4.35 | 0.74 | 0.00 | 1.02 | 1.08 | 1.02 | 3.74 | 0.52 |
| | Proteobacteria | 32.70 | 29.81 | 23.53 | 18.75 | 10.87 | 10.29 | 11.35 | 19.69 | 7.22 | 15.99 | 11.23 | 9.28 |
| | Chlamydiae | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.51 | 0.00 | 0.34 | 0.00 | 0.00 |
| | Cyanobacteria | 0.00 | 0.00 | 1.18 | 0.31 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.53 | 0.00 |
| | Total | 70.14 | 74.53 | 74.71 | 68.13 | 75.00 | 85.29 | 74.67 | 76.47 | 77.98 | 73.81 | 81.28 | 73.71 |
| Eukaryota | Ascomycota | 4.58 | 1.24 | 0.59 | 5.94 | 8.70 | 2.94 | 3.93 | 8.44 | 1.44 | 6.46 | 4.81 | 3.61 |
| | Streptophyta | 1.58 | 2.48 | 0.59 | 1.25 | 3.26 | 1.47 | 0.44 | 0.77 | 1.08 | 3.06 | 0.00 | 2.06 |
| | Chordata | | | | | | | | | | | | |
| | Basidiomycota | 0.00 | 0.00 | 0.00 | 0.31 | 0.00 | 0.00 | 0.44 | 0.77 | 0.00 | 1.02 | 1.60 | 0.52 |
| | Arthropoda | 0.47 | 1.24 | 0.00 | 0.63 | 0.00 | 0.00 | 0.44 | 0.00 | 1.08 | 0.34 | 0.00 | 3.09 |
| | Total | 6.64 | 4.97 | 1.18 | 8.13 | 11.96 | 4.41 | 5.24 | 9.97 | 3.61 | 10.88 | 6.42 | 9.28 |
| | Unassigned | 21.80 | 19.88 | 22.94 | 21.25 | 13.04 | 6.62 | 17.03 | 12.53 | 16.97 | 13.95 | 11.23 | 13.92 |
| | Other | 1.26 | 0.62 | 0.59 | 1.56 | 0.00 | 3.68 | 2.62 | 1.28 | 0.36 | 1.02 | 1.60 | 0.52 |

Annotation revealed no identified rRNA features in the AP-based dataset

Rarefaction analysis showing a number of distinct annotated species as a function of the sequencing reads number was performed on randomly subsampled metagenomic datasets (49,000 reads per sample) annotated against the M5NR non-redundant protein database. The analysis showed the differences in biodiversity (highest level of taxonomic resolution) of the datasets generated by the three metagenomic sequencing approaches (Figure 5.1). The SH- and WGA-based datasets demonstrated similar numbers of identified species from location A and B. A two fold lower number of species were identified in the AP-based dataset. The rarefaction curves computed for all metagenomic datasets did not reach the plateau phase suggesting that more sequencing effort would be required in all methods to achieve species saturation.

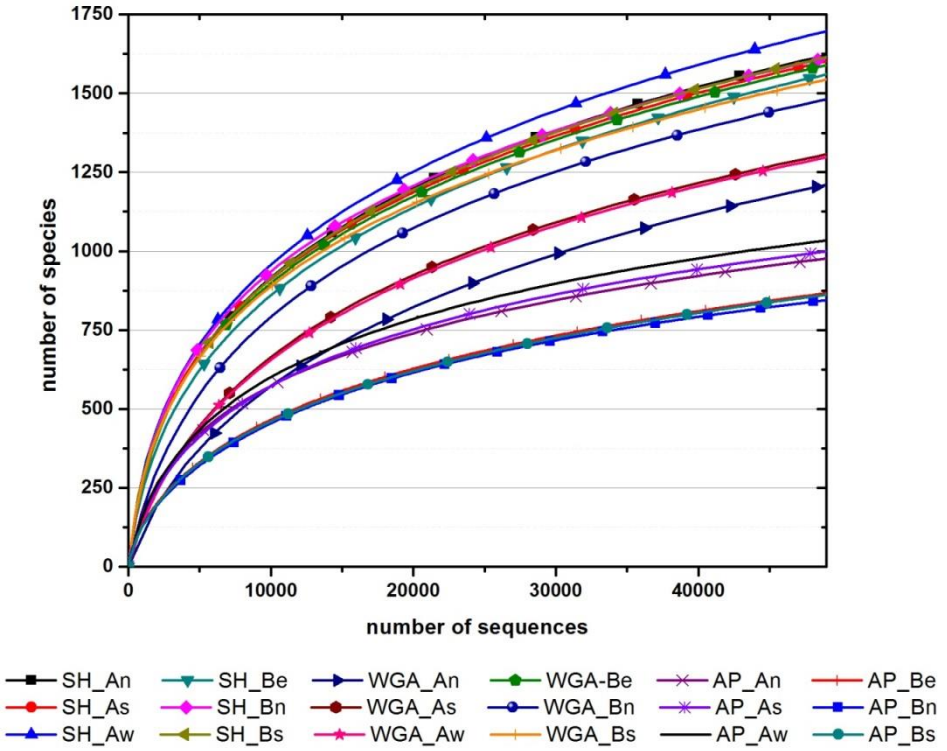



**Figure 5.1. Rarefaction curves created in MG-RAST**. Rarefaction analysis was performed at the species level for each metagenomic protein-derived taxonomic profile based on randomly sub-sampled datasets (49,000 reads per sample). The curves for all taxa include Bacteria, Archaea, Eukaryota, Viruses, unclassified and other sequences identified after metagenomic dataset annotation with M5NR database.

### 5.3.3 Metabolic profiling of metagenomes

Metabolic profiles for all datasets were created by matching to the SEED Subsystems database. The most abundant metabolic features found in all datasets, accounting for almost 60% of assigned reads and the assembled contigs were: clustering-based subsystems; carbohydrates; amino acids and derivatives; protein metabolism; miscellaneous; cofactors; vitamins; prosthetic groups; pigments and DNA metabolism. The relative abundance each of the remaining metabolic features represented less than 5% of reads (Table 5.9 and Table 5.10).

**Table 5.9. Protein-derived metabolic taxonomic composition of the soil microbial communities built on the initial sequencing datasets.** Relative abundances of major taxa (phylum level) derived from assignment of protein gene fragments matched to SEED Subsystems database.

| SEED Subsystems | WGA | | | | | | SH | | | | | | AP | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Level 1 | Aw4 | Aw3 | Aw2 | Bw2 | Bw3 | Bw4 | Aw3 | Aw2 | Aw4 | Bw4 | Bw2 | Bw3 | Aw2 | Aw4 | Aw3 | Bw3 | Bw2 | Bw4 |
| Iron acquisition and metabolism | 0.9 | 0.9 | 1.1 | 0.9 | 0.8 | 0.7 | 0.8 | 0.6 | 0.6 | 0.7 | 1.0 | 0.7 | 0.1 | 0.2 | 0.2 | 0.5 | 0.4 | 0.7 |
| Secondary Metabolism | 0.6 | 0.6 | 0.5 | 0.7 | 0.5 | 0.5 | 0.4 | 0.6 | 0.6 | 0.5 | 0.5 | 0.5 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 |
| Protein Metabolism | 7.9 | 7.6 | 8.1 | 7.7 | 7.5 | 7.4 | 8.7 | 8.5 | 8.2 | 8.6 | 8.5 | 8.7 | 6.9 | 9.4 | 7.1 | 5.0 | 3.9 | 4.7 |
| Carbohydrates | 12.1 | 12.2 | 11.9 | 12.0 | 12.0 | 13.2 | 12.2 | 12.5 | 12.8 | 12.7 | 11.6 | 12.2 | 18.0 | 19.5 | 18.1 | 15.0 | 13.6 | 16.2 |
| Cofactors, Vitamins, Prosthetic Groups, Pigments | 5.6 | 6.0 | 6.4 | 6.2 | 6.7 | 6.8 | 6.2 | 6.4 | 6.4 | 6.4 | 6.3 | 6.3 | 7.9 | 5.7 | 11.5 | 18.3 | 20.3 | 16.0 |
| Membrane Transport | 3.4 | 3.6 | 3.3 | 2.9 | 3.0 | 3.0 | 3.3 | 3.1 | 3.2 | 3.2 | 3.3 | 3.2 | 1.3 | 4.1 | 2.1 | 0.9 | 0.8 | 1.0 |
| Virulence, Disease and Defense | 2.7 | 2.9 | 2.8 | 3.6 | 2.7 | 2.5 | 2.6 | 2.3 | 2.4 | 2.5 | 2.9 | 2.7 | 0.7 | 2.3 | 0.9 | 0.3 | 0.4 | 0.6 |
| Potassium metabolism | 0.5 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.3 | 0.3 | 0.4 | 0.4 | 0.0 | 0.2 | 0.0 | 0.0 | 0.1 | 0.0 |
| Phosphorus Metabolism | 0.9 | 0.8 | 0.8 | 0.8 | 1.0 | 1.0 | 0.9 | 1.0 | 1.0 | 0.9 | 0.9 | 0.9 | 0.0 | 0.0 | 0.0 | 0.1 | 0.1 | 0.1 |
| Nucleosides and Nucleotides | 2.6 | 2.7 | 2.9 | 3.0 | 3.1 | 3.0 | 2.9 | 3.0 | 2.9 | 2.8 | 2.9 | 2.8 | 3.6 | 4.6 | 3.4 | 3.6 | 9.0 | 6.7 |
| Motility and Chemotaxis | 0.7 | 0.8 | 0.8 | 0.8 | 0.9 | 0.8 | 1.0 | 0.9 | 0.9 | 0.9 | 1.0 | 1.0 | 0.8 | 0.3 | 0.1 | 0.1 | 0.1 | 0.4 |
| Miscellaneous | 8.0 | 7.9 | 7.7 | 7.8 | 7.8 | 7.8 | 7.4 | 7.5 | 7.5 | 7.2 | 7.4 | 7.7 | 9.5 | 7.1 | 17.0 | 17.5 | 15.8 | 17.0 |
| Cell Wall and Capsule | 3.5 | 3.5 | 3.5 | 3.8 | 3.2 | 3.0 | 3.3 | 3.1 | 3.0 | 3.2 | 3.4 | 3.1 | 2.6 | 2.6 | 1.2 | 1.2 | 0.5 | 0.9 |
| Phages, Prophages, Transposable elements, Plasmids | 3.4 | 2.9 | 2.2 | 2.2 | 1.3 | 1.4 | 1.5 | 1.6 | 1.7 | 1.5 | 1.4 | 1.5 | 0.5 | 0.6 | 0.3 | 0.2 | 0.2 | 0.4 |
| Fatty Acids, Lipids, and Isoprenoids | 2.5 | 2.8 | 2.8 | 2.8 | 3.2 | 3.2 | 2.9 | 3.0 | 3.0 | 3.0 | 3.0 | 2.9 | 1.3 | 2.5 | 1.1 | 0.4 | 0.3 | 0.4 |
| Dormancy and Sporulation | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| RNA Metabolism | 3.9 | 3.8 | 4.1 | 3.8 | 3.5 | 3.1 | 3.9 | 3.8 | 3.8 | 3.6 | 3.9 | 3.7 | 5.4 | 3.4 | 5.8 | 4.5 | 2.8 | 4.0 |
| Respiration | 3.3 | 3.0 | 3.3 | 3.5 | 3.5 | 3.9 | 3.7 | 3.7 | 3.8 | 3.9 | 3.5 | 3.8 | 4.6 | 7.6 | 3.4 | 0.9 | 0.9 | 0.9 |
| Regulation and Cell signaling | 1.6 | 1.5 | 1.6 | 1.7 | 1.4 | 1.3 | 1.3 | 1.4 | 1.3 | 1.3 | 1.4 | 1.3 | 0.4 | 1.2 | 0.5 | 0.3 | 0.2 | 0.4 |
| Amino Acids and Derivatives | 7.6 | 7.8 | 7.9 | 7.9 | 9.0 | 9.1 | 8.5 | 8.5 | 8.6 | 8.5 | 8.5 | 8.5 | 10.4 | 7.5 | 6.4 | 8.4 | 5.6 | 6.4 |
| DNA Metabolism | 4.8 | 4.9 | 4.4 | 4.7 | 4.2 | 4.0 | 4.1 | 4.2 | 4.1 | 4.1 | 4.4 | 4.3 | 7.6 | 4.2 | 1.2 | 1.8 | 1.4 | 1.3 |
| Nitrogen Metabolism | 0.9 | 1.0 | 1.0 | 1.1 | 1.4 | 1.5 | 1.2 | 1.3 | 1.3 | 1.5 | 1.2 | 1.3 | 0.8 | 1.5 | 0.9 | 0.6 | 0.3 | 0.8 |
| Photosynthesis | 0.1 | 0.1 | 0.2 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Metabolism of Aromatic Compounds | 1.7 | 1.6 | 1.7 | 1.8 | 2.1 | 2.1 | 1.9 | 1.9 | 1.9 | 2.0 | 1.9 | 1.8 | 1.4 | 1.9 | 0.6 | 0.2 | 0.2 | 0.5 |
| Sulfur Metabolism | 1.5 | 1.6 | 1.4 | 1.4 | 1.3 | 1.3 | 1.3 | 1.3 | 1.5 | 1.4 | 1.2 | 1.3 | 2.2 | 2.0 | 3.0 | 1.5 | 1.8 | 1.7 |
| Stress Response | 2.8 | 2.8 | 2.6 | 2.3 | 2.8 | 2.8 | 2.8 | 2.8 | 2.9 | 2.7 | 2.7 | 2.9 | 2.0 | 1.5 | 1.3 | 1.1 | 0.9 | 0.7 |
| Clustering-based subsystems | 14.6 | 14.5 | 15.0 | 14.9 | 14.8 | 14.7 | 15.2 | 15.0 | 15.0 | 15.1 | 15.1 | 14.8 | 11.5 | 9.9 | 13.5 | 17.3 | 16.7 | 16.5 |
| Cell Division and Cell Cycle | 1.5 | 1.6 | 1.5 | 1.3 | 1.4 | 1.1 | 1.3 | 1.3 | 1.3 | 1.3 | 1.3 | 1.3 | 0.3 | 0.3 | 0.2 | 0.2 | 3.9 | 1.9 |

**Table 5.10. Protein-derived metabolic taxonomic composition of the soil microbial communities built on the assembled contigs. Relative abundances of major taxa (phylum level) derived from assignment of protein gene fragments matched to SEED Subsystems database.**

| Metabolic features, SEED subsystem Level 1 | WGA | | | | | | SH | | | | | | AP | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Aw4 | Aw3 | Aw2 | Bw2 | Bw3 | Bw4 | Aw3 | Aw2 | Aw4 | Bw4 | Bw2 | Bw3 | Aw2 | Aw4 | Aw3 | Bw3 | Bw2 | Bw4 |
| Iron acquisition and metabolism | 0.4 | 0.7 | 0.8 | 0.7 | 0.9 | 0.5 | 0.9 | 0.6 | 0.7 | 0.7 | 1.2 | 0.7 | 0.8 | 0.3 | 0.5 | 0.7 | 0.5 | 0.4 |
| Secondary Metabolism | 0.7 | 0.7 | 0.8 | 1.2 | 0.5 | 0.5 | 0.4 | 0.6 | 0.5 | 0.5 | 0.3 | 0.5 | 0.3 | 0.5 | 0.1 | 0.0 | 0.3 | 0.7 |
| Protein Metabolism | 6.0 | 6.8 | 6.0 | 6.6 | 7.4 | 7.8 | 8.9 | 8.6 | 8.2 | 8.6 | 8.2 | 9.0 | 7.0 | 9.2 | 7.2 | 6.1 | 8.4 | 9.0 |
| Carbohydrates | 10.4 | 9.8 | 10.2 | 10.6 | 11.6 | 13.9 | 12.1 | 12.9 | 12.3 | 14.6 | 12.1 | 12.5 | 12.3 | 13.1 | 12.0 | 8.1 | 10.4 | 10.4 |
| Cofactors, Vitamins, Prosthetic Groups, Pigments | 5.5 | 5.4 | 5.6 | 5.5 | 6.5 | 6.3 | 5.9 | 5.6 | 6.8 | 5.7 | 6.3 | 6.0 | 8.2 | 6.2 | 8.2 | 6.1 | 14.2 | 5.2 |
| Membrane Transport | 4.7 | 4.7 | 3.9 | 2.9 | 3.8 | 3.1 | 3.9 | 3.3 | 3.3 | 3.2 | 3.7 | 2.8 | 3.0 | 3.3 | 4.0 | 2.7 | 2.7 | 3.7 |
| Virulence, Disease and Defense | 2.9 | 3.7 | 3.3 | 3.2 | 2.6 | 2.5 | 2.6 | 2.6 | 2.2 | 2.4 | 2.8 | 2.9 | 1.6 | 3.2 | 1.7 | 1.4 | 1.6 | 1.5 |
| Potassium metabolism | 0.5 | 0.6 | 0.3 | 0.4 | 0.5 | 0.5 | 0.5 | 0.3 | 0.4 | 0.6 | 0.6 | 0.2 | 0.1 | 0.3 | 0.1 | 0.0 | 0.8 | 0.4 |
| Phosphorus Metabolism | 1.1 | 1.3 | 0.9 | 0.8 | 1.0 | 1.1 | 0.8 | 1.1 | 0.9 | 0.8 | 1.1 | 0.9 | 0.1 | 0.2 | 0.3 | 0.0 | 0.3 | 0.7 |
| Nucleosides and Nucleotides | 2.1 | 1.4 | 2.3 | 2.6 | 2.1 | 2.9 | 2.8 | 2.3 | 2.2 | 2.4 | 2.6 | 2.5 | 3.3 | 4.4 | 3.0 | 6.8 | 4.6 | 5.6 |
| Motility and Chemotaxis | 0.8 | 1.3 | 1.1 | 1.0 | 1.0 | 0.6 | 1.2 | 1.2 | 0.8 | 1.0 | 1.0 | 1.0 | 0.8 | 0.8 | 0.5 | 0.7 | 0.5 | 0.4 |
| Miscellaneous | 7.5 | 7.1 | 7.6 | 7.5 | 8.3 | 8.1 | 7.2 | 8.1 | 8.3 | 7.2 | 7.6 | 7.5 | 9.4 | 6.7 | 10.2 | 6.8 | 11.7 | 9.7 |
| Cell Wall and Capsule | 3.1 | 2.7 | 3.2 | 3.4 | 3.2 | 3.0 | 3.0 | 3.3 | 3.0 | 3.1 | 3.6 | 3.5 | 2.9 | 3.6 | 3.6 | 4.1 | 2.2 | 4.1 |
| Phages, Prophages, Transposable elements, Plasmids | 7.8 | 7.4 | 5.1 | 4.6 | 1.8 | 1.5 | 1.6 | 1.6 | 2.0 | 1.8 | 1.3 | 1.9 | 1.2 | 1.0 | 0.9 | 2.0 | 0.8 | 1.1 |
| Fatty Acids, Lipids, and Isoprenoids | 1.9 | 3.0 | 2.9 | 2.7 | 3.8 | 3.2 | 3.3 | 3.0 | 2.8 | 3.4 | 3.2 | 2.7 | 1.8 | 3.8 | 2.6 | 1.4 | 0.5 | 0.7 |
| Dormancy and Sporulation | 0.2 | 0.1 | 0.2 | 0.2 | 0.2 | 0.2 | 0.1 | 0.1 | 0.2 | 0.2 | 0.3 | 0.2 | 0.0 | 0.1 | 0.1 | 0.0 | 0.0 | 0.0 |
| RNA Metabolism | 3.9 | 3.3 | 4.2 | 3.9 | 3.0 | 3.0 | 4.4 | 4.6 | 4.3 | 3.7 | 3.8 | 4.5 | 4.7 | 3.1 | 5.1 | 4.1 | 4.6 | 4.1 |
| Respiration | 2.3 | 2.5 | 3.3 | 4.3 | 3.3 | 4.0 | 3.8 | 3.9 | 3.4 | 4.5 | 4.0 | 4.1 | 2.9 | 3.9 | 2.6 | 4.1 | 2.5 | 4.1 |
| Regulation and Cell signaling | 2.8 | 2.0 | 3.1 | 2.5 | 1.5 | 1.5 | 1.4 | 1.4 | 1.4 | 1.1 | 1.7 | 1.1 | 0.8 | 1.1 | 1.5 | 0.7 | 0.5 | 1.1 |
| Amino Acids and Derivatives | 4.8 | 6.2 | 6.8 | 6.6 | 8.0 | 8.6 | 7.8 | 7.1 | 8.1 | 7.6 | 8.1 | 7.8 | 11.9 | 9.6 | 10.8 | 13.5 | 10.4 | 13.8 |
| DNA Metabolism | 6.6 | 6.0 | 4.3 | 4.9 | 4.2 | 3.5 | 3.8 | 3.9 | 3.7 | 3.3 | 4.0 | 3.8 | 5.2 | 5.5 | 3.8 | 7.4 | 4.9 | 5.2 |
| Nitrogen Metabolism | 1.0 | 1.0 | 0.9 | 1.0 | 1.6 | 1.4 | 1.2 | 1.0 | 1.1 | 1.3 | 1.1 | 1.6 | 1.4 | 1.5 | 1.4 | 0.7 | 0.5 | 0.7 |
| Photosynthesis | 0.2 | 0.1 | 0.2 | 0.1 | 0.2 | 0.2 | 0.1 | 0.1 | 0.1 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Metabolism of Aromatic Compounds | 1.8 | 1.9 | 1.7 | 1.8 | 2.3 | 2.0 | 2.1 | 1.9 | 2.1 | 2.1 | 1.8 | 2.0 | 1.7 | 2.0 | 1.7 | 1.4 | 0.5 | 2.6 |
| Sulfur Metabolism | 1.7 | 1.4 | 1.2 | 1.9 | 1.2 | 1.2 | 1.6 | 1.4 | 1.5 | 1.5 | 1.5 | 1.3 | 1.2 | 1.7 | 1.3 | 1.4 | 1.4 | 1.9 |
| Stress Response | 3.1 | 2.8 | 3.4 | 2.0 | 3.1 | 3.0 | 2.6 | 2.8 | 3.2 | 2.8 | 2.7 | 3.3 | 2.4 | 2.3 | 2.1 | 4.1 | 2.5 | 2.2 |
| Clustering-based subsystems | 14.3 | 14.3 | 15.1 | 15.2 | 15.3 | 15.1 | 15.0 | 15.3 | 15.1 | 14.7 | 14.4 | 14.3 | 14.7 | 12.0 | 14.2 | 13.5 | 12.3 | 9.7 |
| Cell Division and Cell Cycle | 2.1 | 2.3 | 1.8 | 1.8 | 1.2 | 1.2 | 1.1 | 1.4 | 1.2 | 1.2 | 1.1 | 1.3 | 0.4 | 0.6 | 0.6 | 2.7 | 0.3 | 0.7 |

### 5.3.4 Comparison of soil metagenomic profiles based on full sequence datasets.

Previous reports have indicated that comparison of metagenomes at high levels of taxonomic and metabolic classification, i.e. analysis based on more broadly defined categories, results in a more conservative estimate of the distances between metagenomic profiles (Fierer et al. 2012). While comparative analysis of the metagenomic profiles at low levels of taxonomic or functional classification (species or subsystems functions, respectively) shows less overlap between the latters and therefore also frequently used for metagenomic profile discrimination (Jeffries et al. 2011; Håvelsrud et al. 2012). The results of the metagenomic datasets comparison in the current study are presented at all MG-RAST taxonomic (Phylum to species) and metabolic (level 1 to functions) levels of classification.

**Comparison of protein-derived taxonomic profiles.**

An initial comparison of the taxonomic structures of the metagenomes using lowest (coarsest) resolution profiles derived at the Phylum level of taxonomy was performed. CLUSTER analysis with group-average linking based on Bray-Curtis similarity matrices delineated two distinct clusters with a similarity score of 85% formed by samples from AP-based dataset grouped according to the sites from where the samples were taken (Figure 5.2A). These clusters were supported by the SIMPROF analysis that showed statistically significant ($p<0.05$) evidence of genuine clustering, as indicated by red dashed branches on the dendrogram (Fig. 2A). Two samples from WGA_A group that exhibited a similarity of 94% also formed such a cluster. Other samples form SH- and WGA-based datasets formed mixed clusters. For example, a

sample from the WGA_B group formed a united cluster with a sample from the SH_A group and two samples from the SH_B group (similarity 94%). One further cluster consisted of two samples from SH_A and SH_B groups with a similarity of 96%.

Pair-wise Bray-Curtis similarity scores between metagenomic profiles were then used to demonstrate inter-samples relationship illustrated on an NMDS plot (Figure 5.2B). It is evident from these figures, NMDS analysis did not reveal a clear visual separation of data. Points denoting samples from WGA- and SH-based datasets were located much closer together showing a higher similarity of the profiles than points representing AP-based dataset (Figure 5.2B). Overlaying clusters on the NMDS plot improved visualisation of the patterns formed by AP-based dataset (Figure 5.2B).

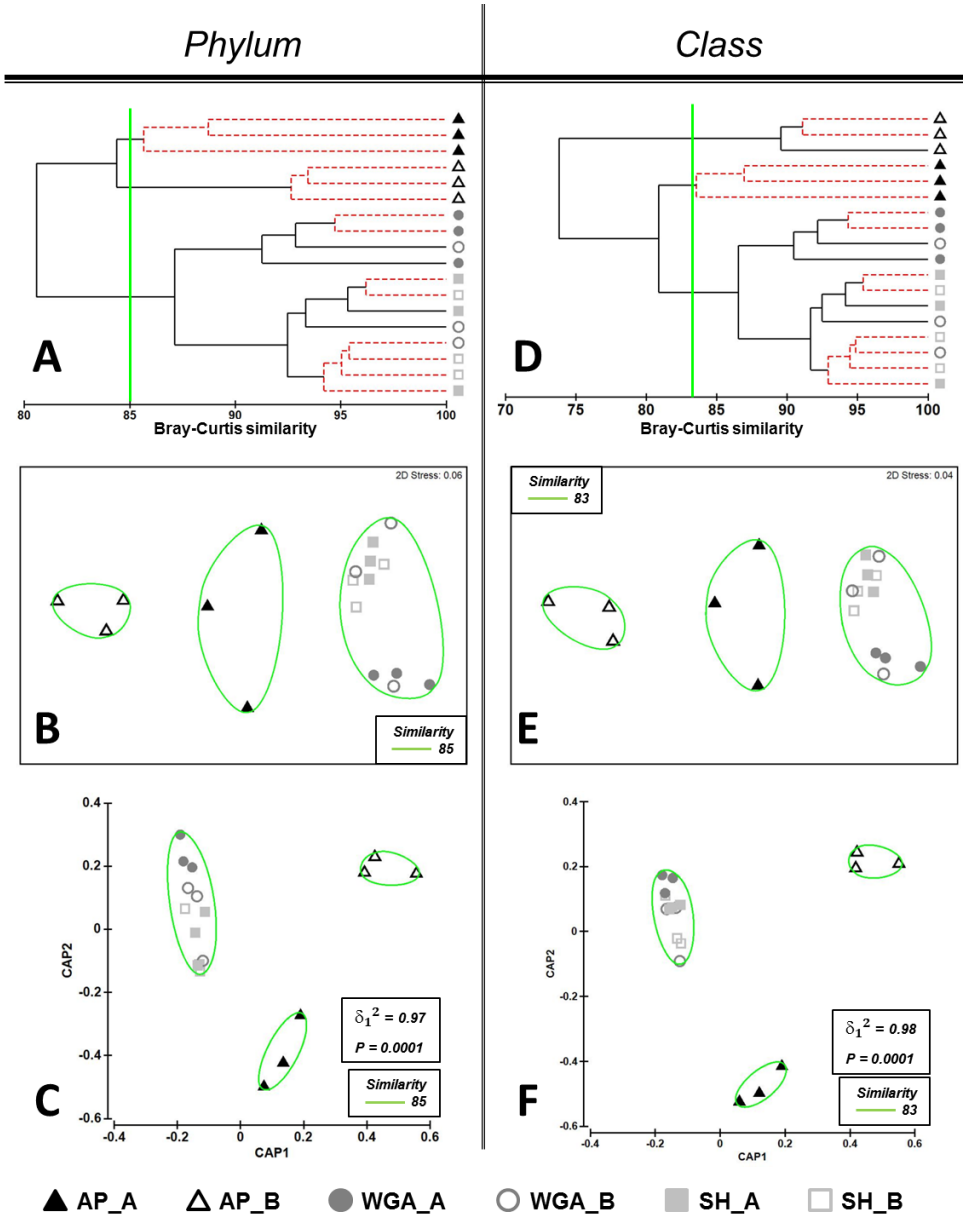



**Figure 5.2. Comparison of the taxonomic soil profiles generated on full datasets at the Phylum (A, B, C) and class (D, E, F) resolution levels.** Bray-Curtis distance similarity matrix was calculated from the square-root transformed abundance of DNA fragments matching taxa in the M5NR database (E-value < $1\times10^{-5}$). The Bray-Curtis matrix was used for generating CLUSTER dendrogram, NMDS and CAP ordination plots. **CLUSTER analysis (A and D).** Red dashed branches on the CLUSTER dendrogram indicate no significant difference between metagenomic profiles (supported by the SIMPROF analysis, $p<0.05$). **NMDS unconstrained ordination (B and E).** The NMDS plot displays distances between samples. Data points that are closer to each other represent samples with highly similar metagenomic profiles. **CAP constrained ordination (C and F).** CAP analysis tests for differences among the groups in multivariate space. The significance of group separation along the canonical axis is indicated by the value of the squared canonical correlation ($\delta_1^2$) and P-value. A contour line on the NMDS and CAP ordinations drawn round each of the cluster defines the superimposition of clusters from CLUSTER dendrogram at the selected level of similarity.

It has been noted that the distinct patterns of datasets separation in the multi-dimensional space could be obscured in the low-dimensional space of NMDS ordination (Anderson & Willis 2003). Consequently for the comparison of these metagenomic datasets, CAP analysis as a constrained ordination method was also performed. CAP analysis tests the hypothesis of whether there is a difference between pre-defined groups. In this study, all datasets were divided into 6 groups in accordance with combined factors, including the sequencing approach applied and the origin of soil samples. The results of the CAP ordination at the Phylum level demonstrated that the first squared canonical correlation was very large ($\delta_1^2$= 0.97), indicating the significance of the CAP model. The first canonical axis showed clear separation of the samples within AP-based dataset according to the soil sampling sites. At the same time a close overlap of the samples from the SH- and WGA-based datasets was observed (Figure 5.2C). However, the cross-validation results of the CAP model for the chosen number of principal coordinate axis ($m$ = 6) did not confirm the above defined separation of the metagenomic datasets (Table 5.11). Thus, the most distinct groups, which had a 100% success under cross-validation, were AP_B and WGA_A. One sample from the AP_A group was misclassified to the AP_B group. One sample from each of the SH_A and SH_B groups were misclassified to the SH_B and SH_A groups, respectively. All the samples from the WGA_B group were misclassified to another of the three different groups (SH_A, SH_B and WGA_A).

**Table 5.11. Results of CAP model cross-validation of soil taxonomic profiles discrimination generated from full sequencing datasets.**

| Original Group | AP_A | AP_B | WGA_A | WGA_B | SH_A | SH_B |
|---|---|---|---|---|---|---|
| **Taxonomy level** | *phylum* ($m$ = 6, $\delta_1^2$ = 0.97, P = 0.0001) | | | | | |
| **% correct** | 67 | 100 | 100 | 0 | 67 | 67 |
| **correct/total** | 2/3 | 3/3 | 3/3 | 0/3 | 2/3 | 2/3 |
| **Misclassified to group** | AP_B | n/a | n/a | SH_A SH_B WGA_A | SH_B | SH_A |
| **Taxonomy level** | *class* ($m$ = 5, $\delta_1^2$ = 0.98, P = 0.0001) | | | | | |
| **% correct** | 100 | 100 | 100 | 0 | 67 | 33 |
| **correct/total** | 3/3 | 3/3 | 3/3 | 0/3 | 2/3 | 1/3 |
| **Misclassified to group** | n/a | n/a | n/a | SH_B WGA_A SH_A | WGA_B | SH_A WGA_B |
| **Taxonomy level** | *order* ($m$ = 3, $\delta_1^2$ = 0.97, P = 0.0002) | | | | | |
| **% correct** | 100 | 100 | 100 | 0 | 67 | 33 |
| **correct/total** | 3/3 | 3/3 | 3/3 | 0/3 | 2/3 | 1/3 |
| **Misclassified to group** | n/a | n/a | n/a | SH_B WGA_A SH_A | SH_B | SH_A WGA_B |
| **Taxonomy level** | *family* ($m$ = 10, $\delta_1^2$ = 0.99, P = 0.0034) | | | | | |
| **% correct** | 100 | 100 | 100 | 0 | 67 | 67 |
| **correct/total** | 3/3 | 3/3 | 3/3 | 0/3 | 2/3 | 2/3 |
| **Misclassified to group** | n/a | n/a | n/a | SH_B WGA_A SH_A | SH_B | WGA_B |
| **Taxonomy level** | *genus* ($m$ = 11, $\delta_1^2$ = 0.99, P = 0.01) | | | | | |
| **% correct** | 100 | 100 | 100 | 0 | 100 | 67 |
| **correct/total** | 3/3 | 3/3 | 3/3 | 0/3 | 3/3 | 2/3 |
| **Misclassified to group** | n/a | n/a | n/a | SH_B WGA_A SH_A | n/a | WGA_B |
| **Taxonomy level** | *species* ($m$ = 10, $\delta_1^2$ = 0.99, P = 0.0065) | | | | | |
| **% correct** | 100 | 100 | 100 | 0 | 67 | 67 |
| **correct/total** | 3/3 | 3/3 | 3/3 | 0/3 | 2/3 | 2/3 |
| **Misclassified to group** | n/a | n/a | n/a | SH_B WGA_A SH_A | SH_B | WGA_B |

It is of note that apart from AP_A and WGA_A groups at the Class level of taxonomic resolution, the cross-validation of the CAP model showed a 100% correct classification of the samples from AP_B group (Table 5.11). Additionally one sample from the SH_A group was misclassified to the WGA_B group, whereas two samples from the SH_B group were misclassified to the SH_A and WGA_B groups.

Further CLUSTER analysis, NMDS and CAP ordinations of the metagenomic profiles at higher levels of taxonomy demonstrated similar patterns of differentiation as observed at the Phylum and Class levels (Figure 5.3, Figure 5.4). Thus, at the Order, Family, genus and species levels of resolution, two samples from the WGA_A group and two samples from the SH_A group formed separate genuine clusters on the CLUSTER dendrograms (Figure 5.3A, Figure 5.3D, Figure 5.4A, Figure 5.4D). Two more genuine mixed clusters were observed consisting of the samples from the SH_B and the WGA_B groups. NMDS and CAP ordinations at all levels of resolution clearly displayed three distinct clusters; two clusters consisting of the samples from the AP_A and the AP_B groups and one mixed cluster of samples from all the other groups (Figure 5.3, Figure 5.4). Cross-validation results of the CAP models at all levels of resolution, starting from the Class level, showed an accurate 100% correct classification of samples from the AP-based dataset (Table 5.11). Despite the visual overlap of the SH- and WGA-based data points shown on the ordination plots (Figure 5.3, Figure 5.4), the samples from WGA_A group were classified 100% correctly across all levels of taxonomic resolution (Table 5.11). Of note was that, at the genus level, all samples from the SH_A group were also successfully allocated.

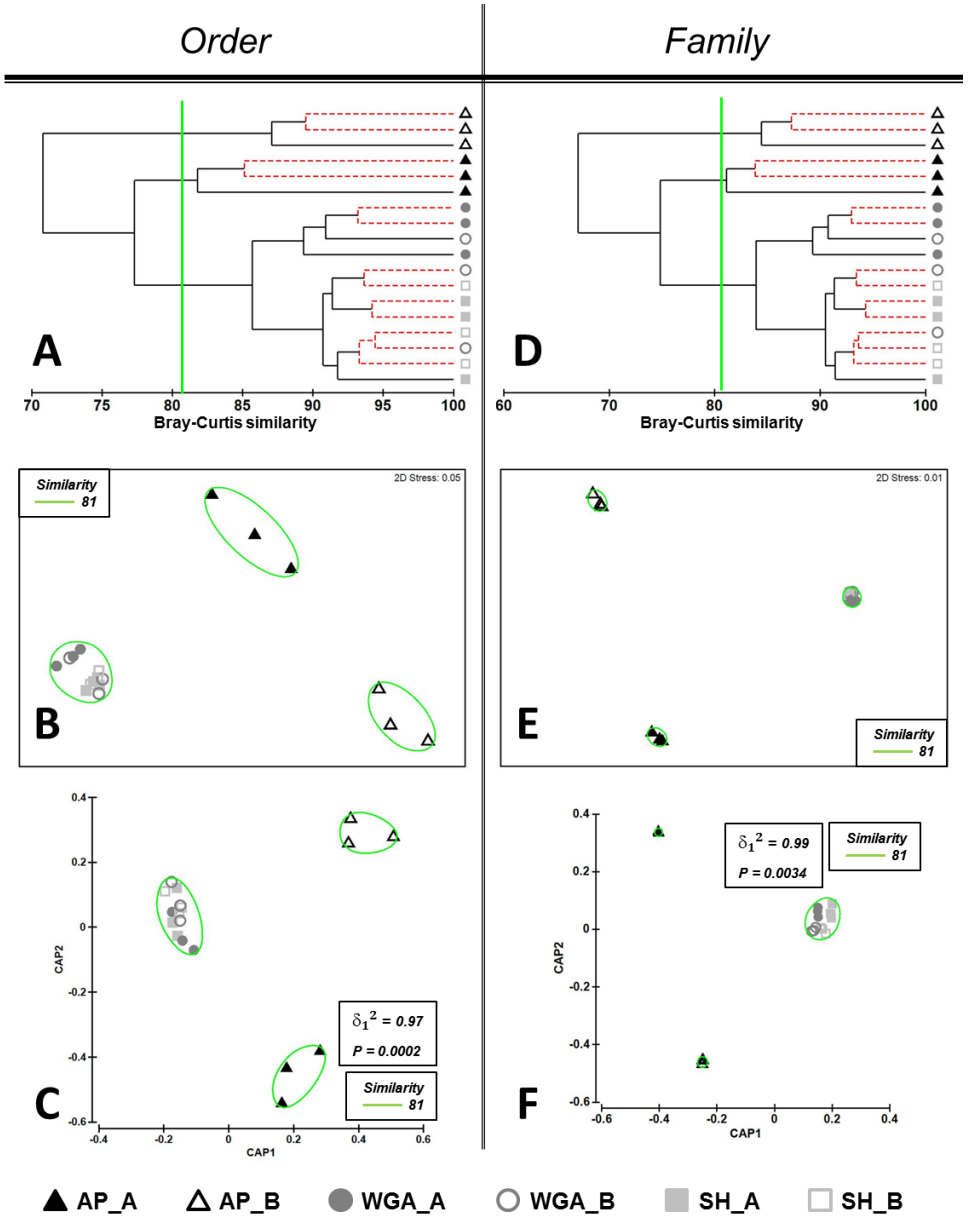



**Figure 5.3. Comparison of the soil protein-derived taxonomic profiles generated on full datasets at the Order (A, B, C) and Family (D, E, F) taxonomic resolution levels.** Bray-Curtis distance similarity matrix was calculated from the square-root transformed abundance of DNA fragments matching taxa to the M5NR database (E-value < 1×10-5). The Bray-Curtis matrix was used for generating CLUSTER dendrogram, NMDS and CAP ordination plots. CLUSTER analysis (A and D). Red dashed branches on the CLUSTER dendrogram indicate no significant difference between metagenomic profiles (supported by the SIMPROF analysis, $p<0.05$). NMDS unconstrained ordination (B and E). The NMDS plot displays distances between samples. Data points that are closer to each other represent samples with highly similar metagenomic profiles. CAP constrained ordination (C and F). CAP analysis tests for differences among the groups in multivariate space. The significance of group separation along the canonical axis is indicated by the value of the squared canonical correlation ($\delta 12$) and P-value ($P<0.05$). A contour line on the NMDS and CAP ordinations drawn round each of the cluster defines the superimposition of clusters from CLUSTER dendrogram at the selected level of similarity.

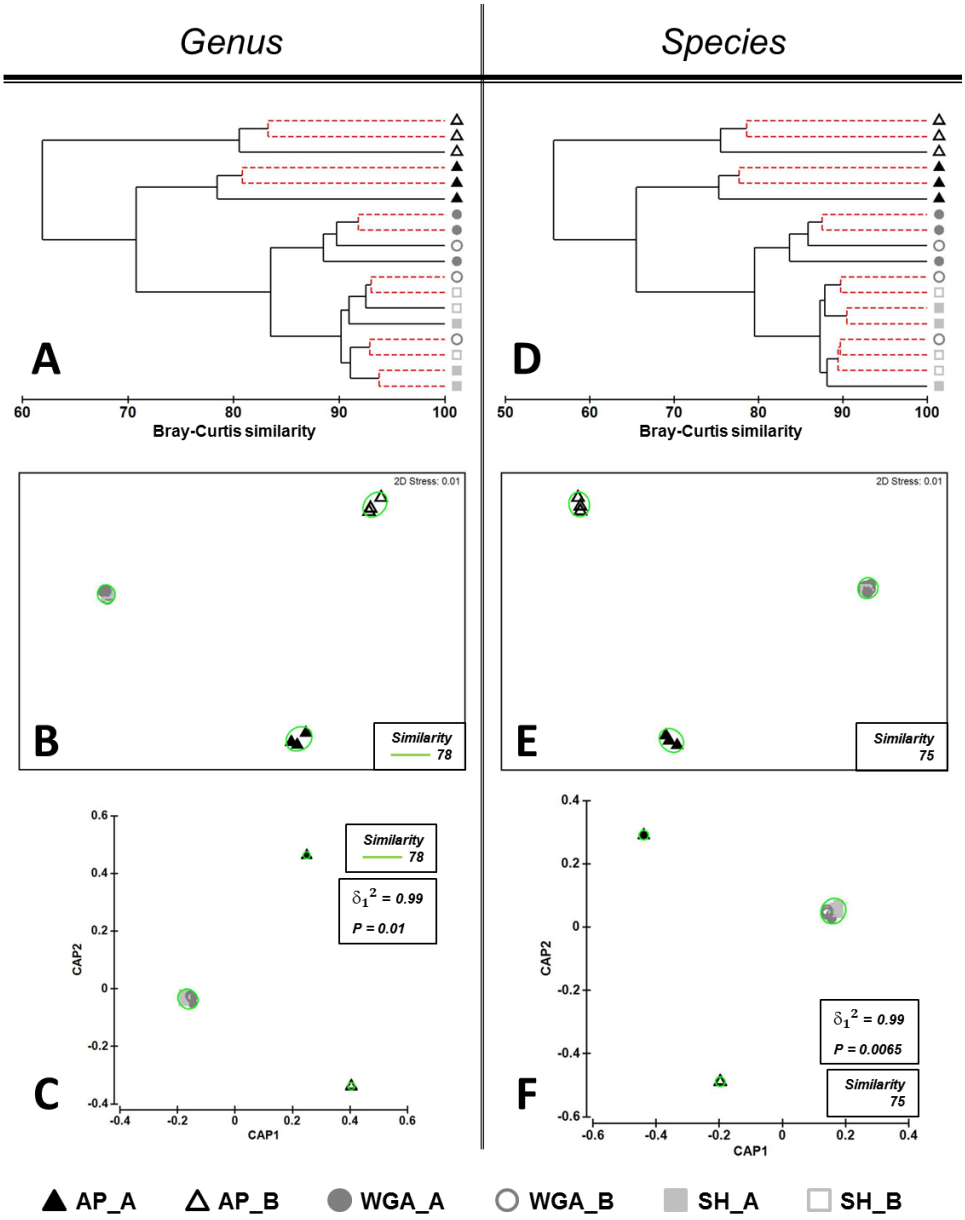



**Figure 5.4. Comparison of the soil protein-derived taxonomic profiles generated on full datasets at the genus (A, B, C) and species (D, E, F) taxonomic resolution levels.** Bray-Curtis distance similarity matrix was calculated from the square-root transformed abundance of DNA fragments matching taxa in the M5NR database (E-value < 1×10-5). The Bray-Curtis matrix was used for generating CLUSTER dendrogram, NMDS and CAP ordination plots. CLUSTER analysis (A and D). Red dashed branches on the CLUSTER dendrogram indicate no significant difference between metagenomic profiles (supported by the SIMPROF analysis, p<0.05). NMDS unconstrained ordination (B and E). The NMDS plot displays distances between samples. Data points that are closer to each other represent samples with highly similar metagenomic profiles. CAP constrained ordination (C and F). CAP analysis tests for differences among the groups in multivariate space. The significance of group separation along the canonical axis is indicated by the value of the squared canonical correlation (δ12) and P-value (P<0.05). A contour line on the NMDS and CAP ordinations drawn round each of the cluster defines the superimposition of clusters from CLUSTER dendrogram at the selected level of similarity.

**Comparison of taxonomic profiles based on rRNA gene fragment classification.**

Taxonomic profiles were generated only for the SH- and the WGA-based datasets where the rRNA gene fragments had matched to data within the M5RNA database. The AP-based dataset was excluded from the consecutive comparative analysis since no sequence matches were found to the ribosomal database. CLUSTER analysis of rRNA-based taxonomic profiles at the Phylum level of resolution demonstrated the formation of four genuine clusters confirmed by SIMPROF analysis ($p<0.05$) (Figure 5.5A). One cluster included three samples from the WGA_A group and one sample from the WGA_B group with a similarity of 77%. A second cluster consisted of two samples from the SH_A group and one sample from the SH_B group with a similarity of 85%. Two other mixed clusters were formed by the samples from different groups. The pattern formed by the samples from the WGA_A group was also observed on the NMDS and CAP plots with a 100% correct allocation; this was confirmed by the results of cross-validation of the CAP model (Figure 5.5B, Figure 5.5C; Table 5.12). Comparison of the metagenomic profiles at higher levels of taxonomic resolution was also performed (Figure 5.5, Figure 5.6, Figure 5.7). Thus, two separate clusters formed by the samples from the WGA_A and the SH_A groups were observed at the genus and species levels (Figure 5.7). Observed groupings had a 100% correct allocation under cross-validation of the CAP model only at the genus level of taxonomic classification (Table 5.12). The latter findings were in full accordance with the allocation of WGA_A and SH_A groups performed using protein-derived taxonomy (Table 5.11).

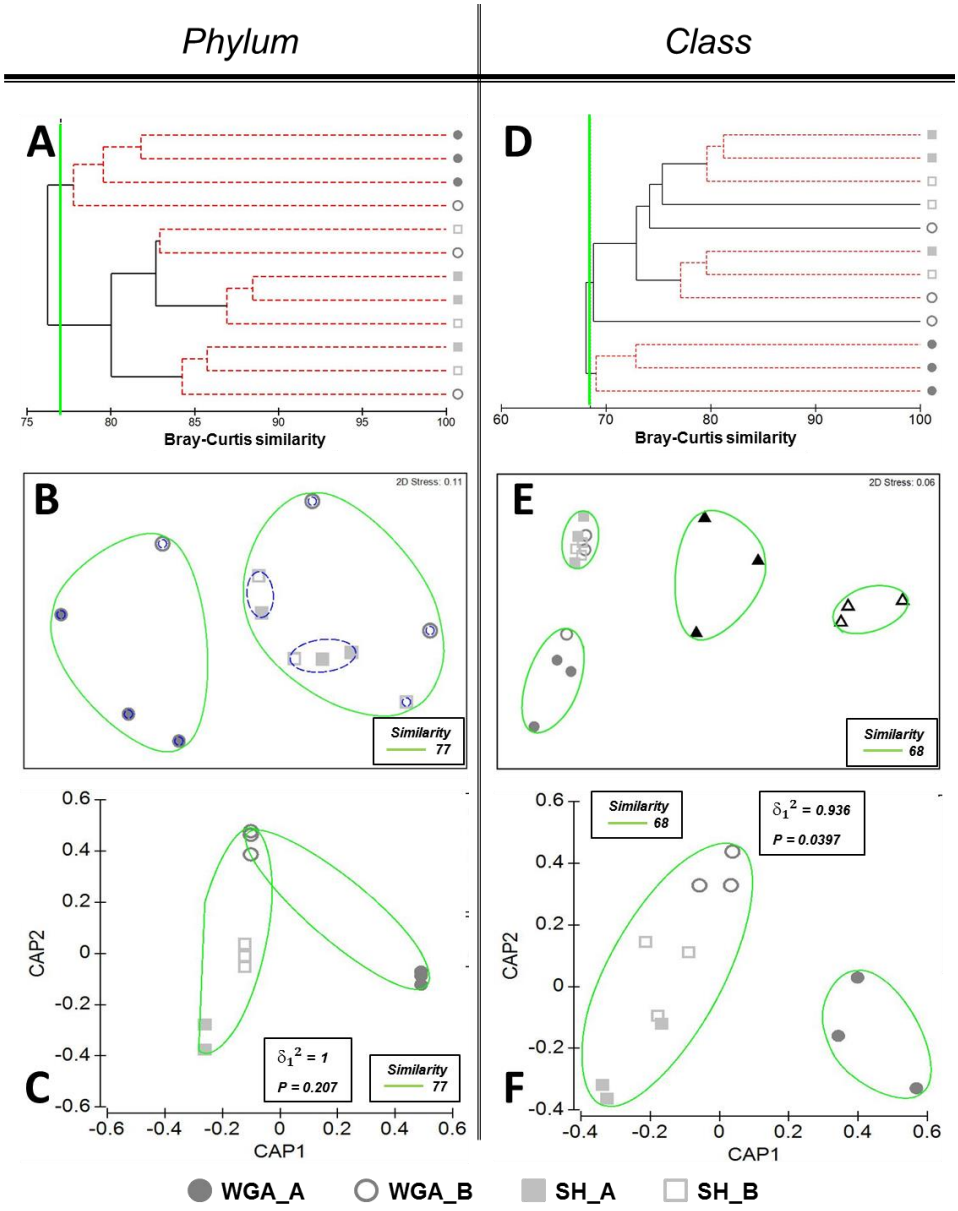



**Figure 5.5. Comparison of the soil rRNA profiles generated on full datasets at the phylum (A, B, C) and class (D, E, F) taxonomic resolution levels.** Bray-Curtis distance similarity matrix was calculated from the square-root transformed abundance of DNA fragments matching taxa in the M5RNA database (E-value < $1\times10^{-5}$). The Bray-Curtis matrix was used for generating CLUSTER dendrogram, NMDS and CAP ordination plots. **CLUSTER analysis (A and D)**. Red dashed branches on the CLUSTER dendrogram indicate no significant difference between metagenomic profiles (supported by the SIMPROF analysis, $p<0.05$). **NMDS unconstrained ordination (B and E)**. The NMDS plot displays distances between samples. Data points that are closer to each other represent samples with highly similar metagenomic profiles. **CAP constrained ordination (C and F)**. CAP analysis tests for differences among the groups in multivariate space. The significance of group separation along the canonical axis is indicated by the value of the squared canonical correlation ($\delta_1^2$) and P-value ($P<0.05$). A contour line on the NMDS and CAP ordinations drawn round each of the cluster defines the superimposition of clusters from CLUSTER dendrogram at the selected level of similarity.

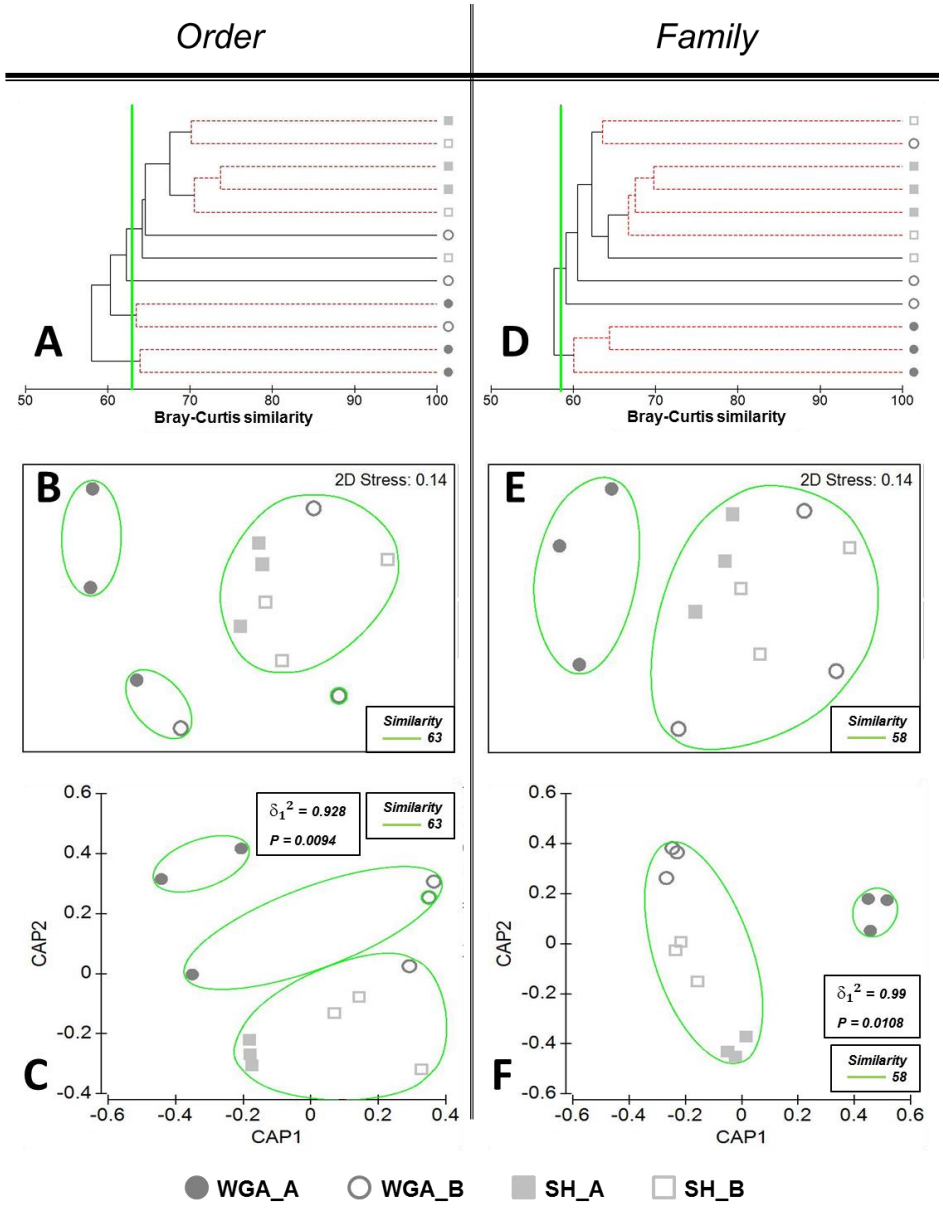



**Figure 5.6. Comparison of the soil rRNA profiles generated on full datasets at the order (A, B, C) and family (D, E, F) taxonomic resolution levels.** Bray-Curtis distance similarity matrix was calculated from the square-root transformed abundance of DNA fragments matching taxa in the M5RNA database (E-value < $1\times10^{-5}$). The Bray-Curtis matrix was used for generating CLUSTER dendrogram, NMDS and CAP ordination plots. **CLUSTER analysis (A and D)**. Red dashed branches on the CLUSTER dendrogram indicate no significant difference between metagenomic profiles (supported by the SIMPROF analysis, $p<0.05$). **NMDS unconstrained ordination (B and E)**. The NMDS plot displays distances between samples. Data points that are closer to each other represent samples with highly similar metagenomic profiles. **CAP constrained ordination (C and F)**. CAP analysis tests for differences among the groups in multivariate space. The significance of group separation along the canonical axis is indicated by the value of the squared canonical correlation ($\delta_1^2$) and P-value ($P<0.05$). A contour line on the NMDS and CAP ordinations drawn round each of the cluster defines the superimposition of clusters from CLUSTER dendrogram at the selected level of similarity.

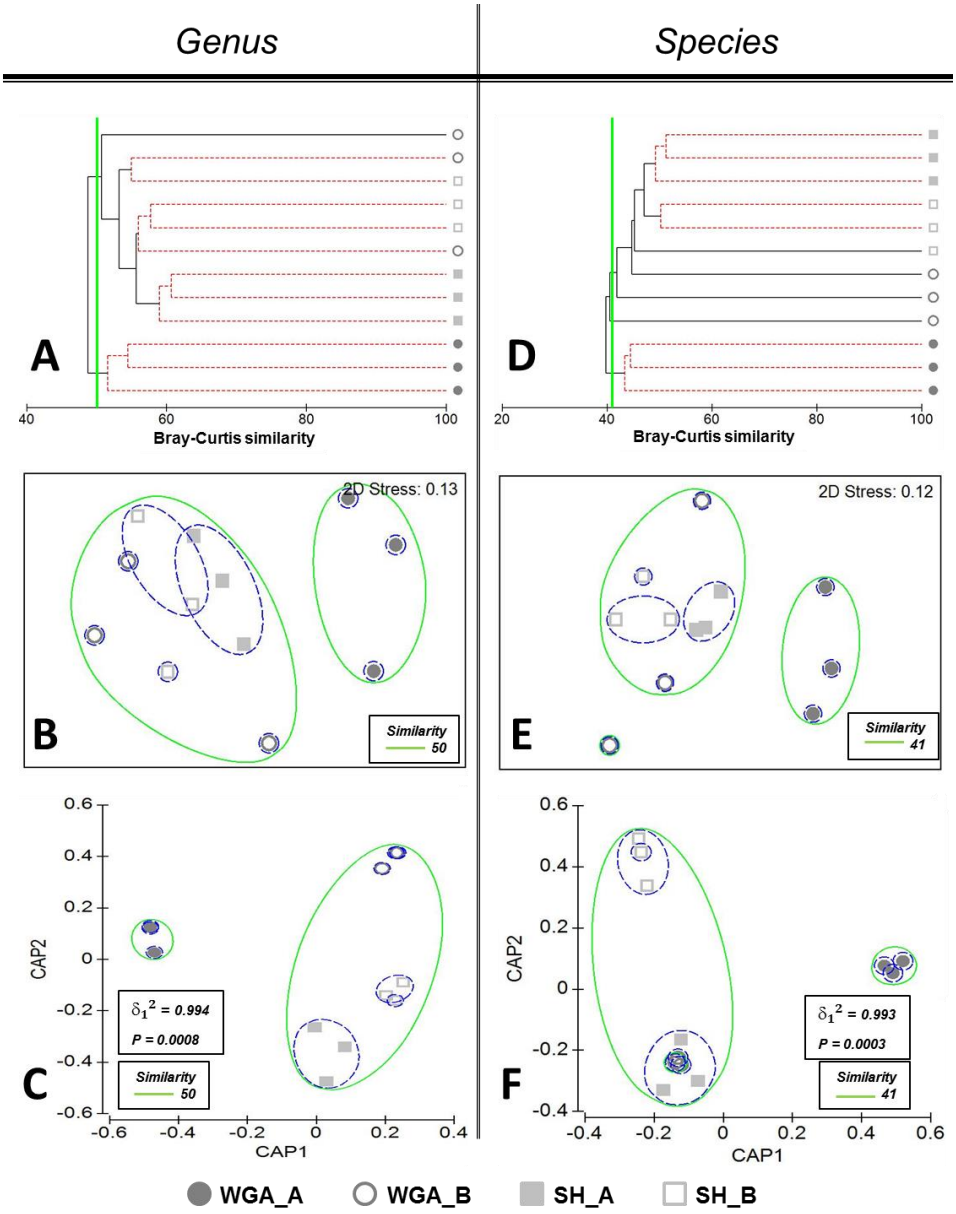



**Figure 5.7. Comparison of the soil rRNA profiles generated on full datasets at the genus (A, B, C) and species (D, E, F) taxonomic resolution levels.** Bray-Curtis distance similarity matrix was calculated from the square-root transformed abundance of DNA fragments matching taxa in the M5RNA database (E-value < $1\times10^{-5}$). The Bray-Curtis matrix was used for generating CLUSTER dendrogram, NMDS and CAP ordination plots. **CLUSTER analysis (A and D)**. Red dashed branches on the CLUSTER dendrogram indicate no significant difference between metagenomic profiles (supported by the SIMPROF analysis, $p<0.05$). **NMDS unconstrained ordination (B and E)**. The NMDS plot displays distances between samples. Data points that are closer to each other represent samples with highly similar metagenomic profiles. **CAP constrained ordination (C and F)**. CAP analysis tests for differences among the groups in multivariate space. The significance of group separation along the canonical axis is indicated by the value of the squared canonical correlation ($\delta_1^2$) and P-value ($P<0.05$). A contour line on the NMDS and CAP ordinations drawn round each of the cluster defines the superimposition of clusters from CLUSTER dendrogram at the selected level of similarity.

**Table 5.12. Results of CAP model cross-validation of soil rRNA taxonomic profiles discrimination generated from full sequencing datasets.**

| Original Group | WGA_A | WGA_B | SH_A | SH_B |
|---|---|---|---|---|
| Taxonomy level | *phylum* ($m$ = 9, $\delta_1^2$ = 1, P = 0.207) | | | |
| % correct | 100 | 33 | 33 | 0 |
| correct/total | 3/3 | 1/3 | 1/3 | 0/3 |
| Misclassified to group | n/a | SH_B | SH_B | SH_A WGA_B |
| Taxonomy level | *class* ($m$ = 5, $\delta_1^2$ = 0.936, P = 0.039) | | | |
| % correct | 33 | 0 | 33 | 100 |
| correct/total | 1/3 | 0/3 | 1/3 | 3/3 |
| Misclassified to group | SH_B | SH_A SH_B | SH_B | n/a |
| Taxonomy level | *order* ($m$ = 4, $\delta_1^2$ = 0.928, P = 0.009) | | | |
| % correct | 66 | 0 | 66 | 100 |
| correct/total | 2/3 | 0/3 | 2/3 | 3/3 |
| Misclassified to group | SH_B | SH_A SH_B | SH_B | n/a |
| Taxonomy level | *family* ($m$ = 6, $\delta_1^2$ = 0.99, P = 0.0108) | | | |
| % correct | 66 | 0 | 66 | 66 |
| correct/total | 2/3 | 0/3 | 2/3 | 2/3 |
| Misclassified to group | SH_B | SH_B | SH_B | WGA_B |
| Taxonomy level | *genus* ($m$ = 8, $\delta_1^2$ = 099, P = 0.0008) | | | |
| % correct | 100 | 0 | 100 | 33 |
| correct/total | 3/3 | 0/3 | 3/3 | 1/3 |
| Misclassified to group | n/a | SH_A SH_B | n/a | SH_A WGA_B |
| Taxonomy level | *species* ($m$ = 8, $\delta_1^2$ = 0.99, P = 0.003) | | | |
| % correct | 100 | 0 | 66 | 33 |
| correct/total | 3/3 | 0/3 | 2/3 | 1/3 |
| Misclassified to group | n/a | SH_B | SH_B | SH_A WGA_B |

**Metabolic profiles comparison.**

CLUSTER analysis of metabolic profiles generated by different sequencing approaches at the lowest level of resolution (level 1) showed that all three samples from the AP_B group formed a separate cluster with a similarity of 92% (Figure 5.8A). Two samples from the AP_A group had a similarity of 90%. The third AP_A sample was bundled with the samples from SH- and WGA- based datasets forming a new mixed cluster. Importantly, the SH- and WGA-based datasets consisting of 12 metagenomic samples formed one united mixed cluster with a similarity of 97% (Figure 5.8A). NMDS and CAP ordinations also showed that all the points associated with the samples from SH- and WGA-based datasets produced a very compact cluster (Figure 5.8B, Figure 5.8C). However, according to a cross-validation procedure, the most distinct groups with 100% allocation success were the AP-based groups and the WGA_A group, whereas misclassification errors were shown for the WGA_B, SH_A and SH_B groups (Table 5.13). Statistical comparisons of the metabolic profiles at higher resolution levels (level 2, level 3 and function) resulted in similar discriminating success (Figure 5.8, Figure 5.9). CLUSTER analysis showed correct site-specific grouping of the samples from AP-based dataset (Figure 5.8D, Figure 5.9A, Figure 5.9D). All the profiles produced by SH- and WGA-based methods again formed a single unresolved cluster. NMDS and CAP ordinations demonstrated clear separation of three clusters (Figure 5.8, Figure 5.9), which was also the case for the metagenomic profiles comparison based on protein-derived taxonomy (Figure 5.2, Figure 5.3, Figure 5.4). In both cases, cross-validation results of the CAP model gave 100% correct classification of the samples from the AP_A, AP_B and WGA_A groups and misclassification errors for samples from the SH_A, SH_B and WGA_B groups (Table 5.11, Table 5.13).

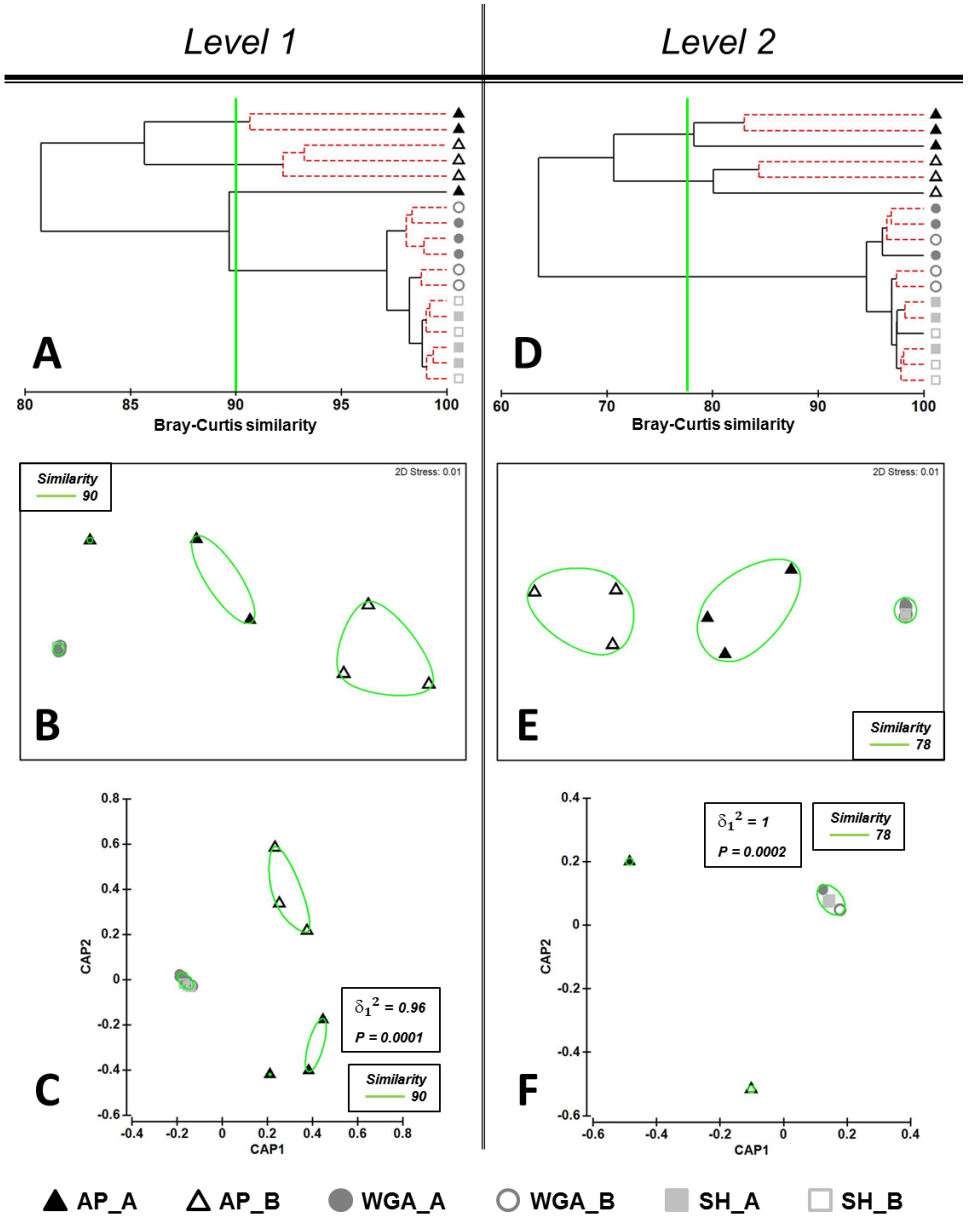



**Figure 5.8. Comparison of the metabolic soil profiles generated on full datasets at the subsystems level 1 (A, B, C) and level 2 (D, E, F) resolution levels.** Bray-Curtis distance similarity matrix was calculated from the square-root transformed abundance of DNA fragments matching taxa in the SEED database (E-value < $1\times10^{-5}$). The Bray-Curtis matrix was used for generating CLUSTER dendrogram, NMDS and CAP ordination plots. **CLUSTER analysis (A and D).** Red dashed branches on the CLUSTER dendrogram indicate no significant difference between metagenomic profiles (supported by the SIMPROF analysis, $p<0.05$). **NMDS unconstrained ordination (B and E).** The NMDS plot displays distances between samples. Data points that are closer to each other represent samples with highly similar metagenomic profiles. **CAP constrained ordination (C and F).** CAP analysis tests for differences among the groups in multivariate space. The significance of group separation along the canonical axis is indicated by the value of the squared canonical correlation ($\delta_1^2$) and P-value. A contour line on the NMDS and CAP ordinations drawn round each of the cluster defines the superimposition of clusters from CLUSTER dendrogram at the selected level of similarity.

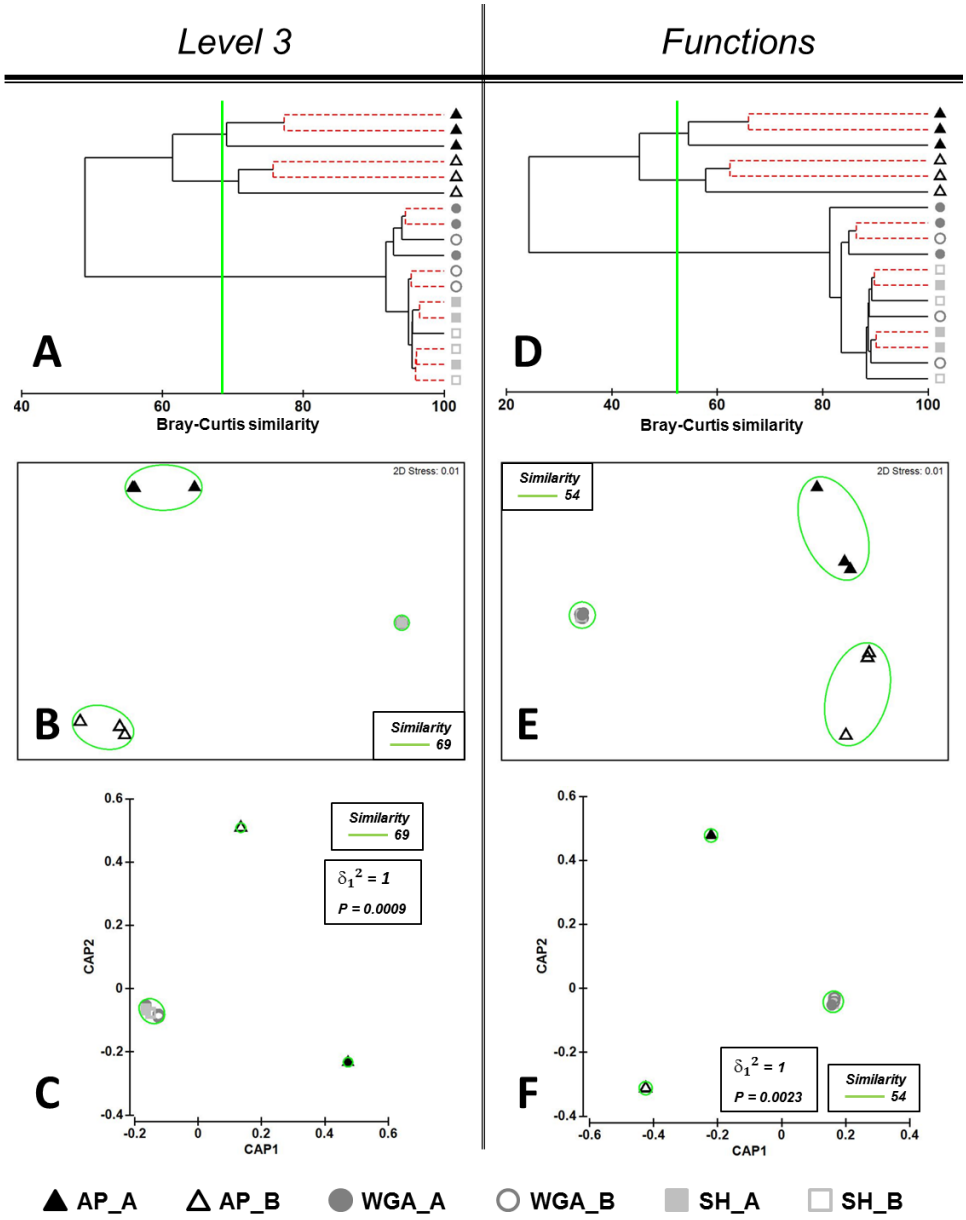



**Figure 5.9. Comparison of the soil metabolic profiles generated on full datasets at the subsystems level 3 (A, B, C) and functions subsystems (D, E, F) metabolic resolution levels.** Bray-Curtis distance similarity matrix was calculated from the square-root transformed abundance of DNA fragments matching taxa in the SEED database (E-value < $1\times10^{-5}$). The Bray-Curtis matrix was used for generating CLUSTER dendrogram, NMDS and CAP ordination plots. **CLUSTER analysis (A and D)**. Red dashed branches on the CLUSTER dendrogram indicate no significant difference between metagenomic profiles (supported by the SIMPROF analysis, p<0.05). **NMDS unconstrained ordination (B and E)**. The NMDS plot displays distances between samples. Data points that are closer to each other represent samples with highly similar metagenomic profiles. **CAP constrained ordination (C and F)**. CAP analysis tests for differences among the groups in multivariate space. The significance of group separation along the canonical axis is indicated by the value of the squared canonical correlation ($\delta_1^2$) and P-value (P<0.05). A contour line on the NMDS and CAP ordinations drawn round each of the cluster defines the superimposition of clusters from CLUSTER dendrogram at the selected level of similarity.

**Table 5.13. Results of CAP model cross-validation of soil metabolic profiles discrimination generated from full sequencing datasets.**

| Original Group | AP_A | AP_B | WGA_A | WGA_B | SH_A | SH_B |
|---|---|---|---|---|---|---|
| Metabolic level | *level 1* ($m$ = 2, $\delta_1^2$ = 0.96, P = 0.0001) | | | | | |
| % correct | 100 | 100 | 100 | 33 | 67 | 33 |
| correct/total | 3/3 | 3/3 | 3/3 | 1/3 | 2/3 | 1/3 |
| Misclassified to group | n/a | n/a | n/a | SH_A SH_B | SH_B | SH_A WGA_B |
| Metabolic level | *level 2* ($m$ = 11, $\delta_1^2$ = 1, P = 0.0002) | | | | | |
| % correct | 100 | 100 | 100 | 33 | 67 | 100 |
| correct/total | 3/3 | 3/3 | 3/3 | 1/3 | 2/3 | 3/3 |
| Misclassified to group | n/a | n/a | n/a | SH_B WGA_A | SH_B | n/a |
| Metabolic level | *level 3* ($m$ = 12, $\delta_1^2$ = 1, P = 0.0009) | | | | | |
| % correct | 100 | 100 | 100 | 33 | 67 | 67 |
| correct/total | 3/3 | 3/3 | 3/3 | 1/3 | 2/3 | 2/3 |
| Misclassified to group | n/a | n/a | n/a | SH_B WGA_A | SH_B | SH_A |
| Metabolic level | *functions* ($m$ = 10, $\delta_1^2$ = 1, P = 0.0023) | | | | | |
| % correct | 100 | 100 | 67 | 0 | 67 | 67 |
| correct/total | 3/3 | 3/3 | 2/3 | 0/3 | 2/3 | 2/3 |
| Misclassified to group | n/a | n/a | WGA_B | SH_B WGA_A SH_B | SH_B | SH_A |

### 5.3.5 Comparison of metagenomic profiles based on randomly sub-sampled datasets.

**Comparison of protein-derived taxonomic and metabolic profiles.**

It has been proposed that in order to enable the comparison of metagenomes based on equal sequencing efforts, the datasets should be randomly sub-sampled to the size of the smallest sample (Gilbert et al. 2010; Fierer et al. 2012). The metagenomic datasets generated by SH-, WGA- and AP-PCR-based approaches were re-analysed by MG-RAST at an equivalent sequencing depth of 49 000 annotated reads per sample. Comparison of taxonomic and metabolic profiles generated from sub-sampled datasets at all levels of classification available within MG-RAST was performed by CLUSTER analysis, NMDS and CAP ordination.

Statistical analysis of the sub-sampled metagenomic datasets generated by three metagenome sequencing approaches yielded nearly identical estimates of the overall differences between soil microbial communities from locations A and B as those obtained using full sequence datasets (Figure 5.10 – Figure 5.14, Table 5.14, Table 5.15). This similarity was also confirmed using the RELATE tool which revealed a strong correlation between Bray-Curtis distance matrices (Spearman rank coefficient $r>0.9$, $p<0.0001$) generated on both full, and sub-sampled, datasets at all levels of taxonomic and metabolic resolution (Table 5.16).

**Comparison of taxonomic profiles based on rRNA gene fragment classification.**

CLUSTER analysis and NMDS ordination of the rRNA-based taxonomic profiles at the Phylum level of taxonomy demonstrated a heterogeneous mixed cluster of the samples from the SH- and WGA-based datasets with an average similarity of

70%. CAP analysis showed 100% correct classification of samples from the WGA_A group and misclassification errors for samples from other groups. At the species level of resolution CLUSTER analysis also revealed a single heterogeneous mixed cluster with the taxonomic profile similarity of approximately 25%. CAP analysis indicated a high degree of misclassification errors.

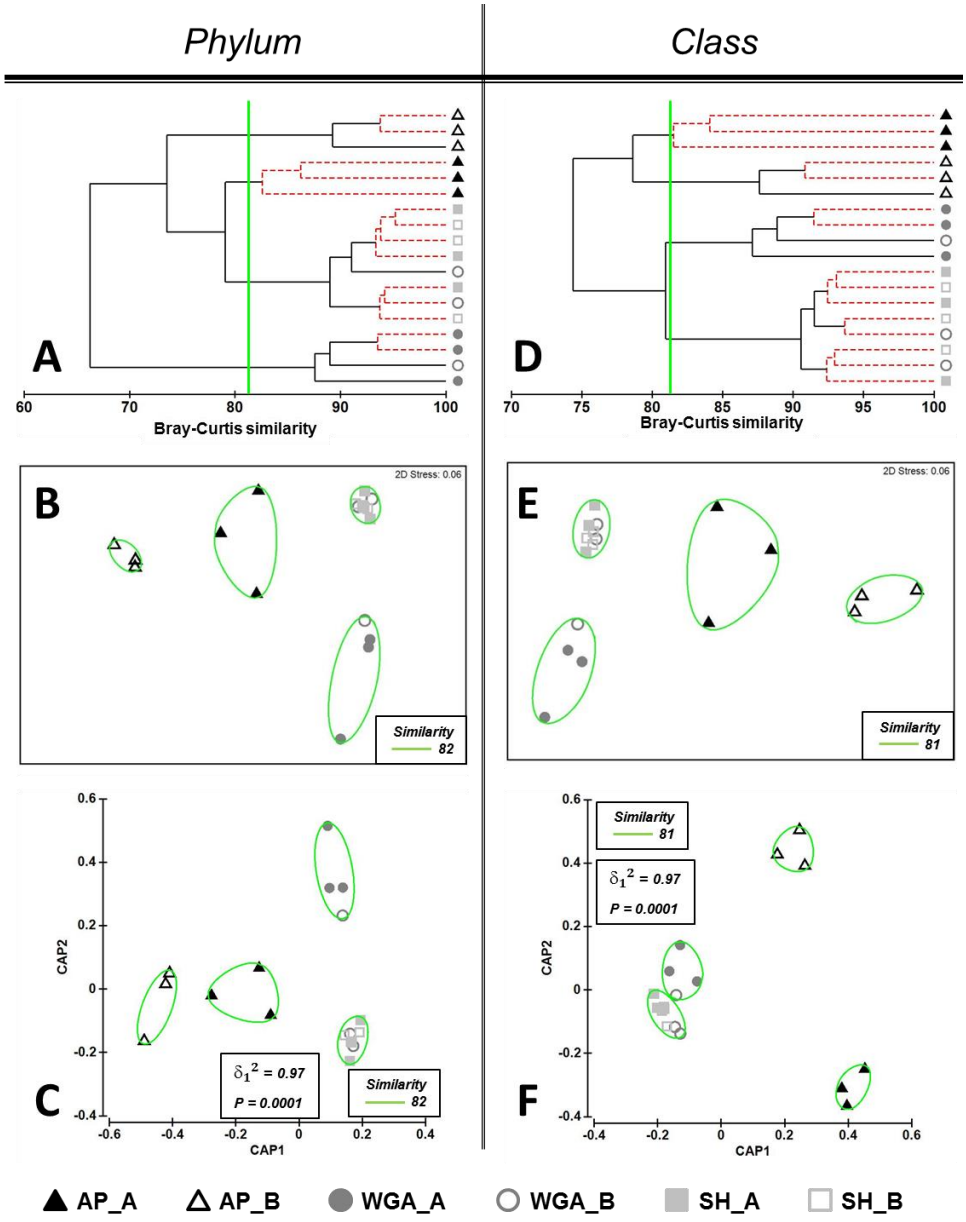



**Figure 5.10. Comparison of the soil protein-derived taxonomic profiles generated on randomly sub-sampled datasets at the phylum (A, B, C) and class (D, E, F) metabolic resolution levels.** Bray-Curtis distance similarity matrix was calculated from the square-root transformed abundance of DNA fragments matching taxa in the M5NR database (E-value < $1\times10^{-5}$). The Bray-Curtis matrix was used for generating CLUSTER dendrogram, NMDS and CAP ordination plots. **CLUSTER analysis (A and D)**. Red dashed branches on the CLUSTER dendrogram indicate no significant difference between metagenomic profiles (supported by the SIMPROF analysis, $p<0.05$). **NMDS unconstrained ordination (B and E)**. The NMDS plot displays distances between samples. Data points that are closer to each other represent samples with highly similar metagenomic profiles. **CAP constrained ordination (C and F)**. CAP analysis tests for differences among the groups in multivariate space. The significance of group separation along the canonical axis is indicated by the value of the squared canonical correlation ($\delta_1^2$) and P-value ($P<0.05$). A contour line on the NMDS and CAP ordinations drawn round each of the cluster defines the superimposition of clusters from CLUSTER dendrogram at the selected level of similarity.

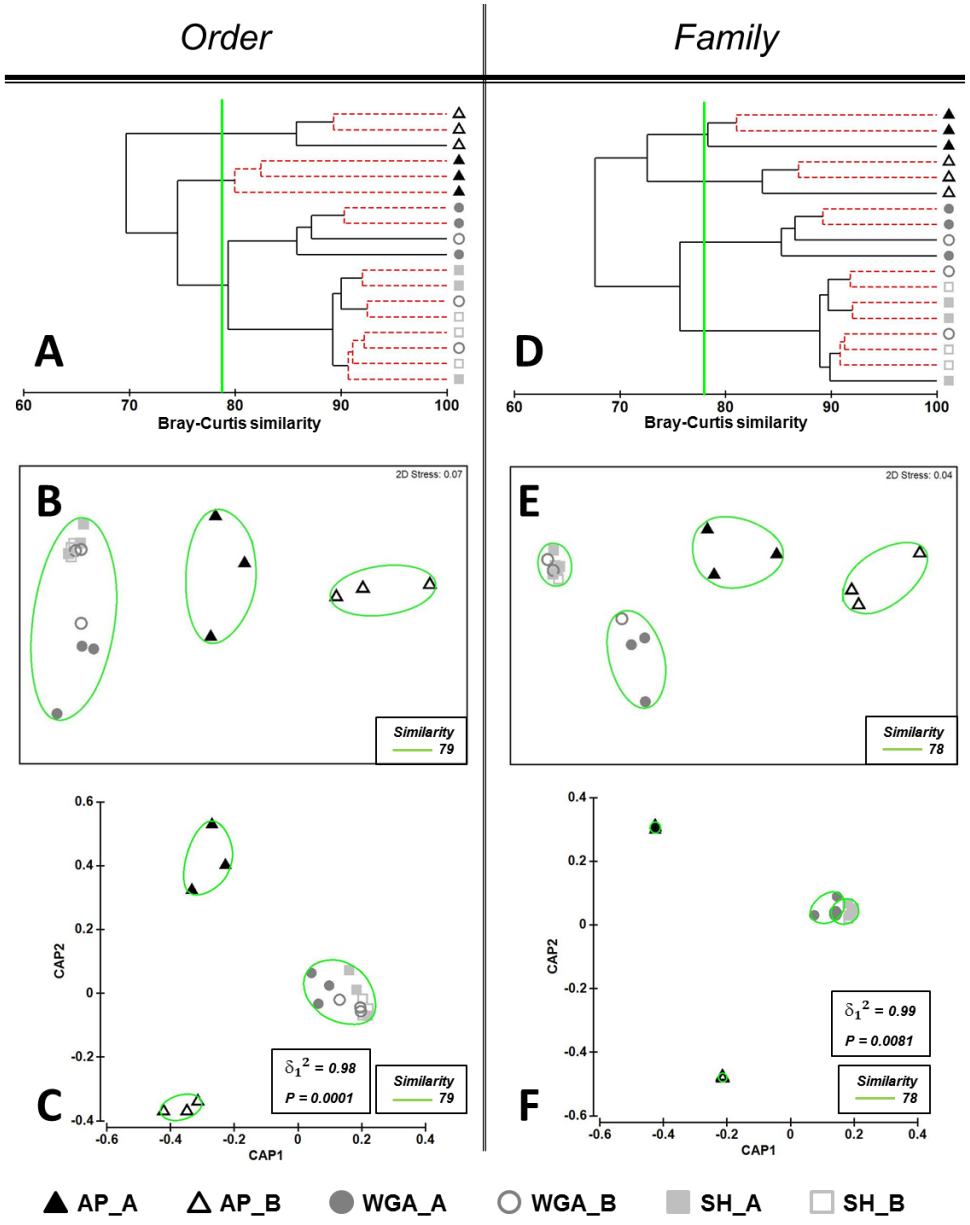



**Figure 5.11. Comparison of the soil protein-derived taxonomic profiles generated on randomly sub-sampled datasets at the order (A, B, C) and family (D, E, F) taxonomic resolution levels.** Bray-Curtis distance similarity matrix was calculated from the square-root transformed abundance of DNA fragments matching taxa in the SEED database (E-value < $1\times10^{-5}$). The Bray-Curtis matrix was used for generating CLUSTER dendrogram, NMDS and CAP ordination plots. **CLUSTER analysis (A and D)**. Red dashed branches on the CLUSTER dendrogram indicate no significant difference between metagenomic profiles (supported by the SIMPROF analysis, $p<0.05$). **NMDS unconstrained ordination (B and E)**. The NMDS plot displays distances between samples. Data points that are closer to each other represent samples with highly similar metagenomic profiles. **CAP constrained ordination (C and F)**. CAP analysis tests for differences among the groups in multivariate space. The significance of group separation along the canonical axis is indicated by the value of the squared canonical correlation ($\delta_1^2$) and P-value ($P<0.05$). A contour line on the NMDS and CAP ordinations drawn round each of the cluster defines the superimposition of clusters from CLUSTER dendrogram at the selected level of similarity.

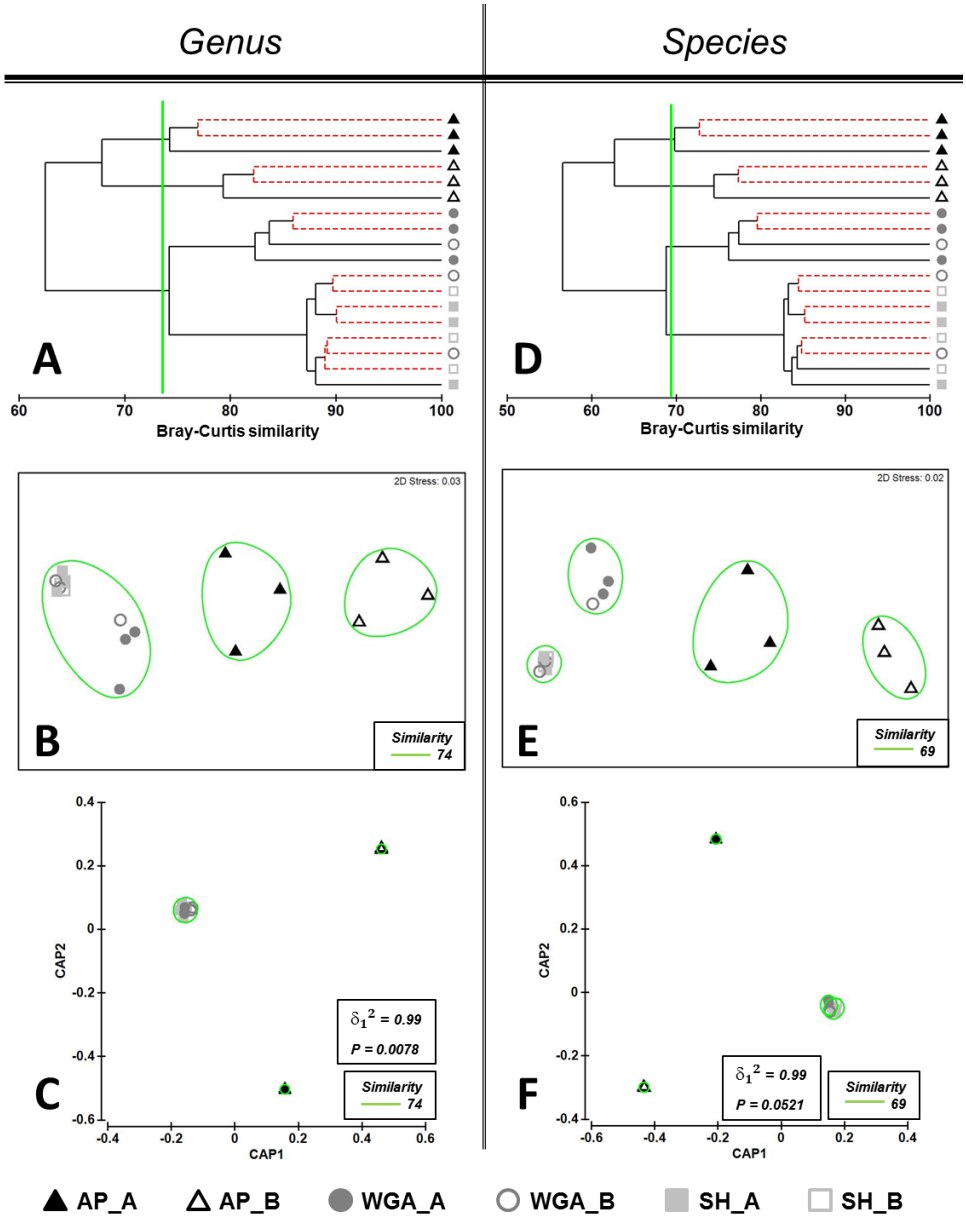



**Figure 5.12. Comparison of the soil protein-derived taxonomic profiles generated on randomly sub-sampled datasets at the genus (A, B, C) and species (D, E, F) taxonomic resolution levels.** Bray-Curtis distance similarity matrix was calculated from the square-root transformed abundance of DNA fragments matching taxa in the M5NR database (E-value < $1\times10^{-5}$). The Bray-Curtis matrix was used for generating CLUSTER dendrogram, NMDS and CAP ordination plots. **CLUSTER analysis (A and D)**. Red dashed branches on the CLUSTER dendrogram indicate no significant difference between metagenomic profiles (supported by the SIMPROF analysis, $p<0.05$). **NMDS unconstrained ordination (B and E)**. The NMDS plot displays distances between samples. Data points that are closer to each other represent samples with highly similar metagenomic profiles. **CAP constrained ordination (C and F)**. CAP analysis tests for differences among the groups in multivariate space. The significance of group separation along the canonical axis is indicated by the value of the squared canonical correlation ($\delta_1^2$) and P-value ($P<0.05$). A contour line on the NMDS and CAP ordinations drawn round each of the cluster defines the superimposition of clusters from CLUSTER dendrogram at the selected level of similarity.

**Table 5.14. Results of CAP model cross-validation of soil protein-derived taxonomic profiles discrimination generated from sub-sampled sequencing datasets.**

| Original Group | AP_A | AP_B | WGA_A | WGA_B | SH_A | SH_B |
|---|---|---|---|---|---|---|
| Taxonomy level | ***Phylum*** ($m$ = 2, $\delta_1^2$ = 0.97, P = 0.0001) | | | | | |
| % correct | 100 | 100 | 100 | 0 | 33 | 33 |
| correct/total | 3/3 | 3/3 | 3/3 | 0/3 | 1/3 | 1/3 |
| Misclassified to group | n/a | n/a | n/a | SH_A SH_B WGA_A | SH_B | SH_A |
| Taxonomy level | ***Class*** ($m$ = 5, $\delta_1^2$ = 0.98, P = 0.0001) | | | | | |
| % correct | 100 | 100 | 100 | 0 | 67 | 67 |
| correct/total | 3/3 | 3/3 | 3/3 | 0/3 | 2/3 | 2/3 |
| Misclassified to group | n/a | n/a | n/a | SH_B WGA_A | SH_B | WGA_B |
| Taxonomy level | ***Order*** ($m$ = 3, $\delta_1^2$ = 0.98, P = 0.0001) | | | | | |
| % correct | 100 | 100 | 100 | 0 | 33 | 67 |
| correct/total | 3/3 | 3/3 | 3/3 | 0/3 | 1/3 | 2/3 |
| Misclassified to group | n/a | n/a | n/a | SH_B WGA_A | SH_B | SH_A |
| Taxonomy level | ***Family*** ($m$ = 9, $\delta_1^2$ = 0.99, P = 0.0081) | | | | | |
| % correct | 100 | 100 | 100 | 0 | 67 | 67 |
| correct/total | 3/3 | 3/3 | 3/3 | 0/3 | 2/3 | 2/3 |
| Misclassified to group | n/a | n/a | n/a | SH_B WGA_A SH_A | SH_B | WGA_B |
| Taxonomy level | ***genus*** ($m$ = 10, $\delta_1^2$ = 0.99, P = 0.0078) | | | | | |
| % correct | 100 | 100 | 100 | 0 | 67 | 67 |
| correct/total | 3/3 | 3/3 | 3/3 | 0/3 | 2/3 | 2/3 |
| Misclassified to group | n/a | n/a | n/a | SH_B WGA_A SH_A | SH_B | SH_A |
| Taxonomy level | ***species*** ($m$ = 10, $\delta_1^2$ = 0.99, P = 0.0621) | | | | | |
| % correct | 100 | 100 | 100 | 0 | 67 | 67 |
| correct/total | 3/3 | 3/3 | 3/3 | 0/3 | 2/3 | 2/3 |
| Misclassified to group | n/a | n/a | n/a | SH_B WGA_A SH_A | SH_B | SH_A |

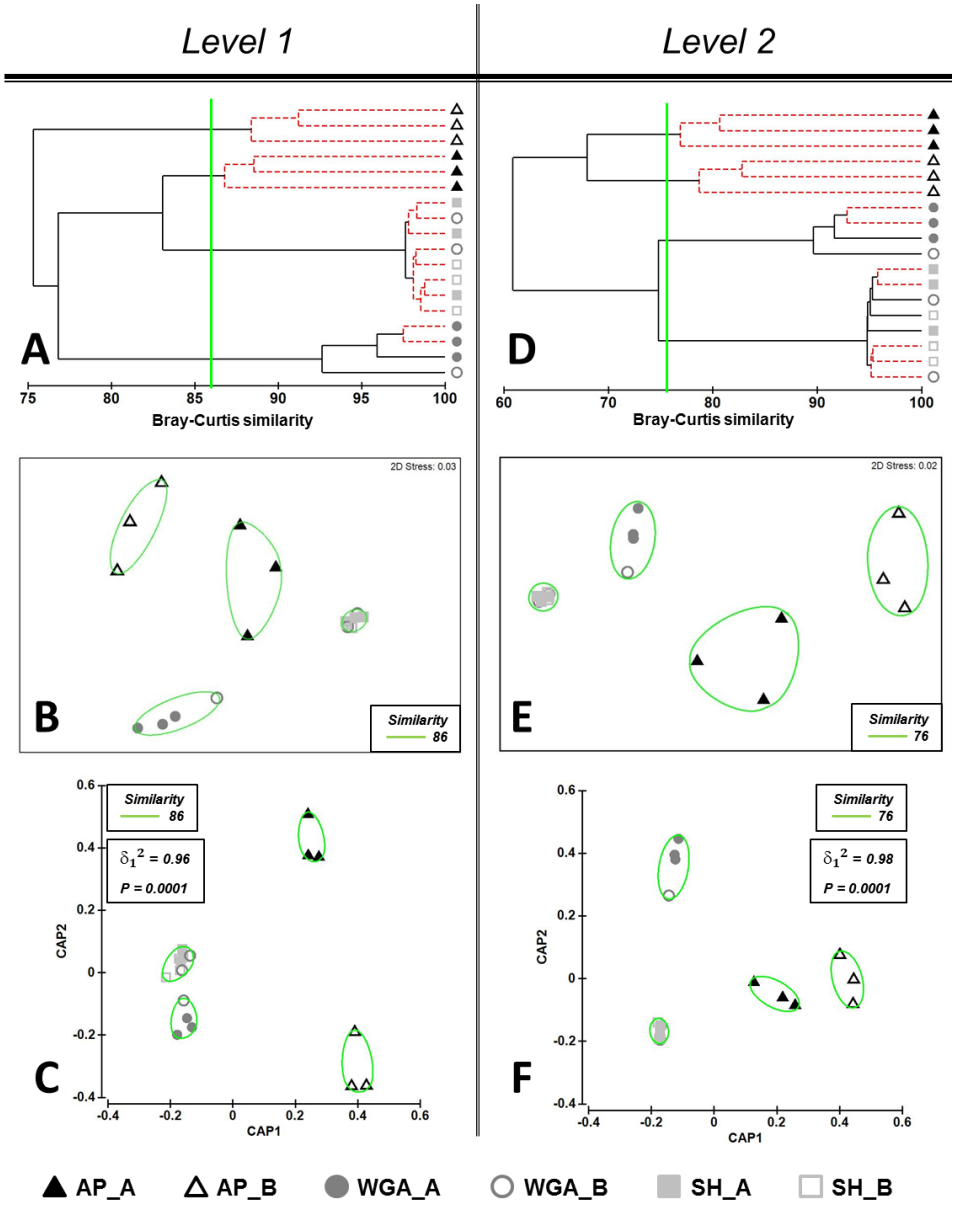



**Figure 5.13. Comparison of the soil metabolic profiles generated on randomly sub-sampled datasets at the subsystems Level 1 (A, B, C) and subsystems Level 2 (D, E, F) metabolic resolution levels.** Bray-Curtis distance similarity matrix was calculated from the square-root transformed abundance of DNA fragments matching taxa in the SEED database (E-value < $1\times10^{-5}$). The Bray-Curtis matrix was used for generating CLUSTER dendrogram, NMDS and CAP ordination plots. **CLUSTER analysis (A and D)**. Red dashed branches on the CLUSTER dendrogram indicate no significant difference between metagenomic profiles (supported by the SIMPROF analysis, p<0.05). **NMDS unconstrained ordination (B and E)**. The NMDS plot displays distances between samples. Data points that are closer to each other represent samples with highly similar metagenomic profiles. **CAP constrained ordination (C and F)**. CAP analysis tests for differences among the groups in multivariate space. The significance of group separation along the canonical axis is indicated by the value of the squared canonical correlation ($\delta_1^2$) and P-value (P<0.05). A contour line on the NMDS and CAP ordinations drawn round each of the cluster defines the superimposition of clusters from CLUSTER dendrogram at the selected level of similarity.

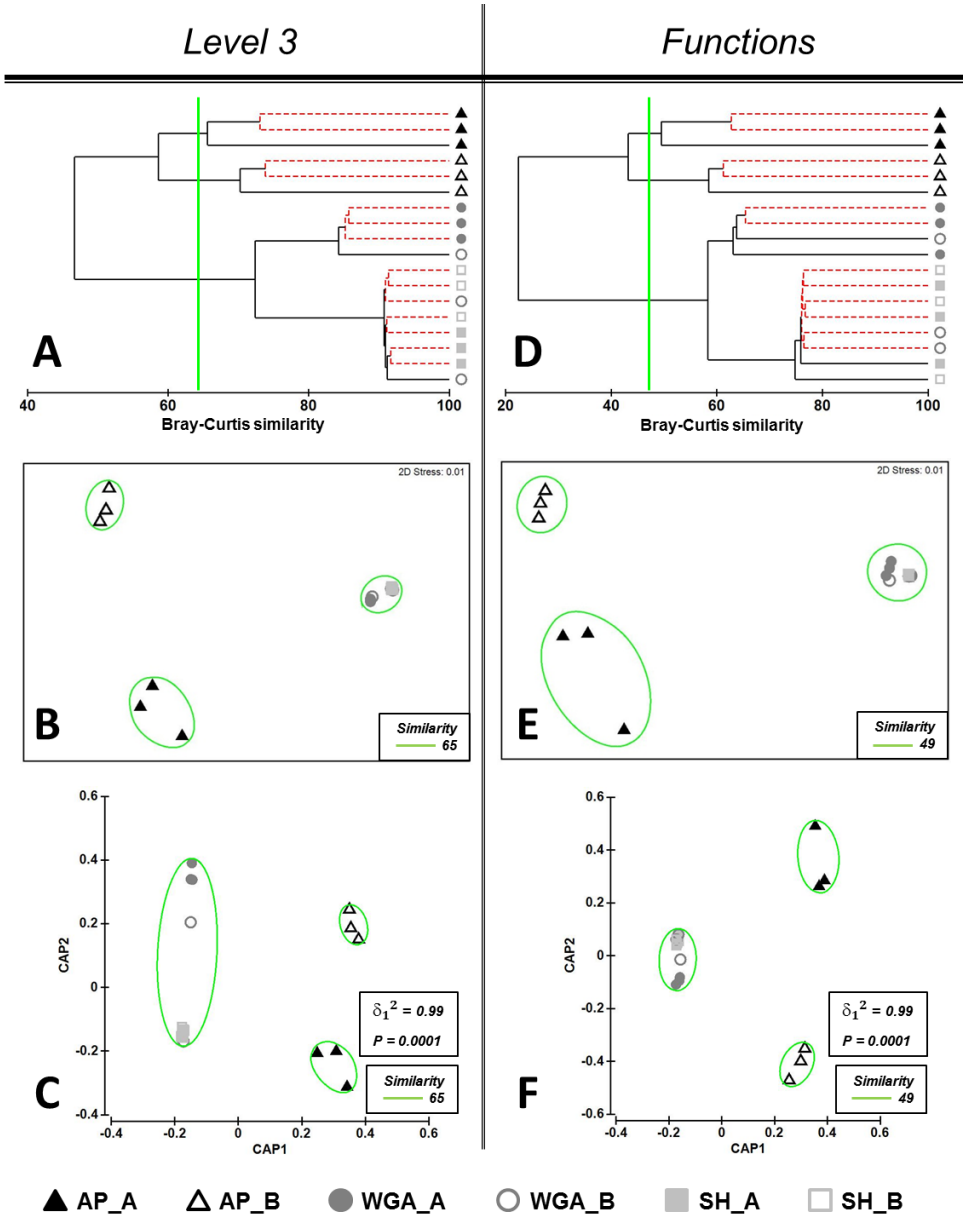



**Figure 5.14. Comparison of the soil metabolic profiles generated on randomly sub-sampled datasets at the subsystems level 3 (A, B, C) and subsystems functions (D, E, F) metabolic resolution levels.** Bray-Curtis distance similarity matrix was calculated from the square-root transformed abundance of DNA fragments matching taxa in the SEED database (E-value < $1\times10^{-5}$). The Bray-Curtis matrix was used for generating CLUSTER dendrogram, NMDS and CAP ordination plots. **CLUSTER analysis (A and D)**. Red dashed branches on the CLUSTER dendrogram indicate no significant difference between metagenomic profiles (supported by the SIMPROF analysis, p<0.05). **NMDS unconstrained ordination (B and E)**. The NMDS plot displays distances between samples. Data points that are closer to each other represent samples with highly similar metagenomic profiles. **CAP constrained ordination (C and F)**. CAP analysis tests for differences among the groups in multivariate space. The significance of group separation along the canonical axis is indicated by the value of the squared canonical correlation ($\delta_1^2$) and P-value (P<0.05). A contour line on the NMDS and CAP ordinations drawn round each of the cluster defines the superimposition of clusters from CLUSTER dendrogram at the selected level of similarity.

**Table 5.15. Results of CAP model cross-validation of soil metabolic profiles discrimination generated from sub-sampled sequencing datasets.**

| Original Group | **AP_A** | **AP_B** | **WGA_A** | **WGA_B** | **SH_A** | **SH_B** |
|---|---|---|---|---|---|---|
| Metabolic level | *level 1* ($m$ = 5, $\delta_1^2$= 0.96, P = 0.0001) | | | | | |
| % correct | 100 | 100 | 100 | 0 | 67 | 67 |
| correct/total | 3/3 | 3/3 | 3/3 | 0/3 | 2/3 | 2/3 |
| Misclassified to group | n/a | n/a | n/a | SH_A SH_B WGA_A | SH_B | SH_A |
| Metabolic level | *level 2* ($m$ = 2, $\delta_1^2$ = 0.99, P = 0.0001) | | | | | |
| % correct | 100 | 100 | 100 | 0 | 67 | 100 |
| correct/total | 3/3 | 3/3 | 3/3 | 0/3 | 2/3 | 3/3 |
| Misclassified to group | n/a | n/a | n/a | SH_B WGA_A SH_A | SH_B | n/a |
| Metabolic level | *level 3* ($m$ = 2, $\delta_1^2$ = 0.99, P = 0.0001) | | | | | |
| % correct | 100 | 100 | 100 | 0 | 67 | 100 |
| correct/total | 3/3 | 3/3 | 3/3 | 0/3 | 2/3 | 3/3 |
| Misclassified to group | n/a | n/a | n/a | SH_B WGA_A SH_A | SH_B | n/a |
| Metabolic level | *functions* ($m$ = 3, $\delta_1^2$ = 0.99, P = 0.0001) | | | | | |
| % correct | 100 | 100 | 100 | 0 | 33 | 100 |
| correct/total | 3/3 | 3/3 | 3/3 | 0/3 | 1/3 | 3/3 |
| Misclassified to group | n/a | n/a | n/a | SH_B WGA_A SH_B | SH_B | n/a |

**Table 5.16. RELATE comparison of Bray-Curtis similarity matrices.**

| Taxonomic level | Spearman rank coefficient | Metabolic level | Spearman rank coefficient |
|---|---|---|---|
| Phylum | 0.887 | level 1 | 0.652 |
| Class | 0.944 | level 2 | 0.958 |
| Order | 0.959 | level 3 | 0.967 |
| Family | 0.940 | functions | 0.969 |
| genus | 0.965 | | |
| species | 0.966 | | |

### 5.3.6 Comparison of soil metagenomic profiles based on the assembled sequence datasets.

Assembly of sequencing reads can serve as a useful tool which can link unassigned reads to those successfully matched to the genomes found in reference database. This can greatly reduce complexity of the metagenomic datasets and, as a result, decrease biases during consecutive annotation process.

In this study assembly of the initial metagenomic datasets was performed using open-source software 'Meta-Velvet' (Namiki et al. 2012). MetaVelvet uses de Bruijn graph-based algorithm of assembly with user defined length of k-mers. The main assembly output is a list of contigs generated a FASTA file. The program is run from the command line and required basic knowledge of the Linux operational system.

The assembled datasets were then annotated by MG-RAST using the same protein and ribosomal reference databases as for the initial datasets. The same multivariate statistical approaches were used for the comparison of taxonomic and metabolic profiles generated from the assembled datasets at all levels of classification available within MG-RAST.

**Comparison of protein-derived taxonomic profiles.**

CLUSTER analysis and NMDS ordination plot of the taxonomic profiles from the assembled datasets at the Phylum and Class levels resulted in formation of mixed clusters consisting of SH, WGA and AP-based datasets (Figure 5.15). CAP analysis demonstrated misclassification errors for samples from all groups (Table 5.17).

CLUSTER analysis and NMDS ordination of the profiles at the Order and Family levels showed the formation of separate clusters made of the profiles from the AP_A and AP_B groups according to the soil sampling sites (Figure 5.16A, B, D, E). CAP ordination plot also displayed distant separation of these clusters as well as clusters formed by the profiles from the SH and WGA groups (Figure 5.16C, Figure 5.16F). However, cross-validation of CAP model at the Order level confirmed 100% correct classification of the profiles from groups AP_A, WGA_A and SH_B only ($\delta_1^2$ = 0.999, p=0 01). Whereas at the Family level, 100% correct classification of the profiles from groups AP_A, AP_B and WGA_A ($\delta_1^2$ = 0.994, p=0 0001) was observed (Table 5.17).

CLUSTER analysis of the profiles generated at the genus and species levels resulted in formation of clear separate cluster consisting of the profiles from group AP_A. Only two profiles from the AP_B group clustered together, while the third one formed an individual neighbour branch (Figure 5.17A, Figure 5.17D). Also only two profiles from the groups WGA_A and SH_A formed genuine small clusters included into the large heterogeneous mixed cluster that also contained the remaining profiles from the SH and WGA groups. It is of note the NMDS ordination plots displayed the profiles from the AP_B, WGA_A, WGA_B, SH_A and SH_B groups as overlaying points indicting their high similarity (Figure 5.17B, Figure 5.17E). However, cross-validation of the CAP model confirmed 100% correct classification of the profiles from

the AP_A and AP_B groups at both genus and species levels of taxonomic resolution (Table 5.17).

**Comparison of metabolic profiles.**

Comparison of the metabolic profiles generated from SH, WGA and AP assembled datasets at the lowest level of resolution (level 1) did not show correct clustering of the samples according their collection sites (Figure 5.18). Cross-validation of the CAP model confirmed 100% correct classification of the samples from the AP_A and WGA_A groups (Table 5.18). The metabolic profiles form AP_A and AP_B groups at the higher levels of resolution, namely at the level 2, level 3 and functions, clustered according to the origin of the corresponding soil samples as shown in the CLUSTER dendrogram, NMDS and CAP plots (Figure 5.18D-F, Figure 5.19D-F). All samples from the SH and WGA groups in turn formed one mixed cluster at the all levels of resolution. Cross-validation of the CAP model confirmed 100% correct classification of the samples from AP_A, AP_B and SH_A groups at all levels of metabolic hierarchy (Table 5.18).

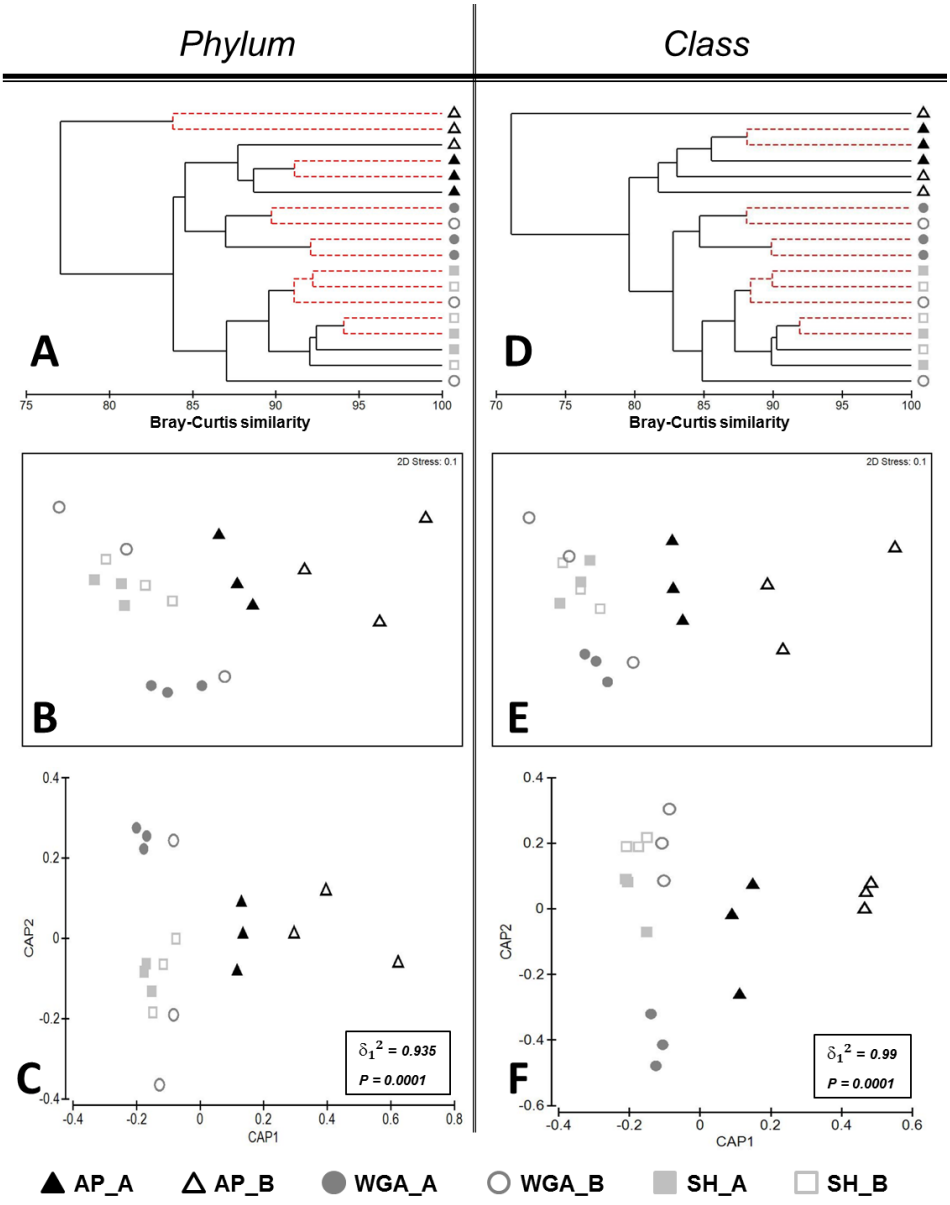



**Figure 5.15. Comparison of the taxonomic soil profiles generated from the assembled contigs at the phylum (A, B, C) and class (D, E, F) resolution levels.** Bray-Curtis distance similarity matrix was calculated from the square-root transformed abundance of DNA fragments matching taxa in the M5NR database (E-value < $1\times10^{-5}$). The Bray-Curtis matrix was used for generating CLUSTER dendrogram, NMDS and CAP ordination plots. **CLUSTER analysis (A and D).** Red dashed branches on the CLUSTER dendrogram indicate no significant difference between metagenomic profiles (supported by the SIMPROF analysis, $p<0.05$). **NMDS unconstrained ordination (B and E).** The NMDS plot displays distances between samples. Data points that are closer to each other represent samples with highly similar metagenomic profiles. **CAP constrained ordination (C and F).** CAP analysis tests for differences among the groups in multivariate space. The significance of group separation along the canonical axis is indicated by the value of the squared canonical correlation ($\delta_1^2$) and P-value. A contour line on the NMDS and CAP ordinations drawn round each of the cluster defines the superimposition of clusters from CLUSTER dendrogram at the selected level of similarity.

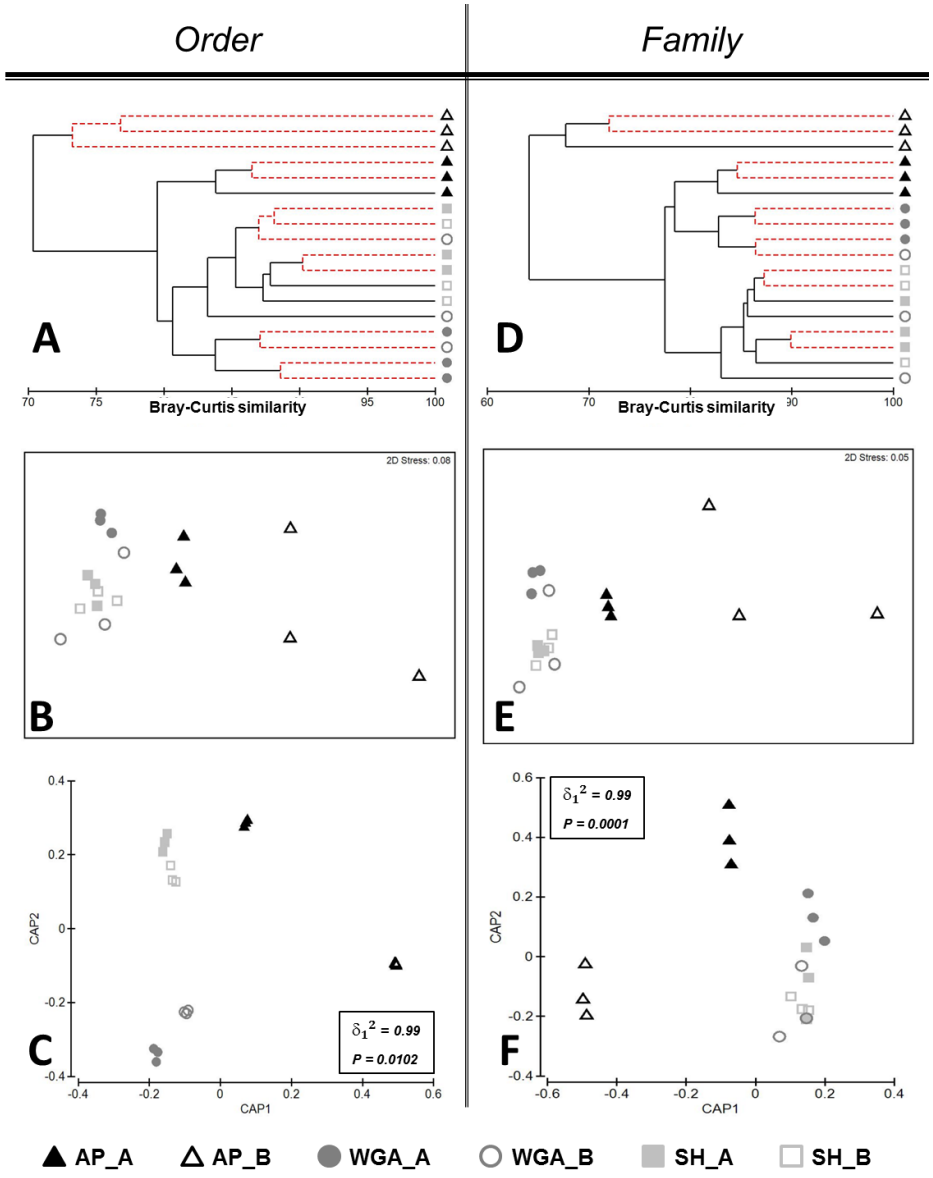



**Figure 5.16. Comparison of the taxonomic soil profiles generated from the assembled contigs at the order (A, B, C) and family (D, E, F) resolution levels.** Bray-Curtis distance similarity matrix was calculated from the square-root transformed abundance of DNA fragments matching taxa in the M5NR database (E-value < $1\times10^{-5}$). The Bray-Curtis matrix was used for generating CLUSTER dendrogram, NMDS and CAP ordination plots. **CLUSTER analysis (A and D).** Red dashed branches on the CLUSTER dendrogram indicate no significant difference between metagenomic profiles (supported by the SIMPROF analysis, $p<0.05$). **NMDS unconstrained ordination (B and E).** The NMDS plot displays distances between samples. Data points that are closer to each other represent samples with highly similar metagenomic profiles. **CAP constrained ordination (C and F).** CAP analysis tests for differences among the groups in multivariate space. The significance of group separation along the canonical axis is indicated by the value of the squared canonical correlation ($\delta_1^2$) and P-value. A contour line on the NMDS and CAP ordinations drawn round each of the cluster defines the superimposition of clusters from CLUSTER dendrogram at the selected level of similarity.

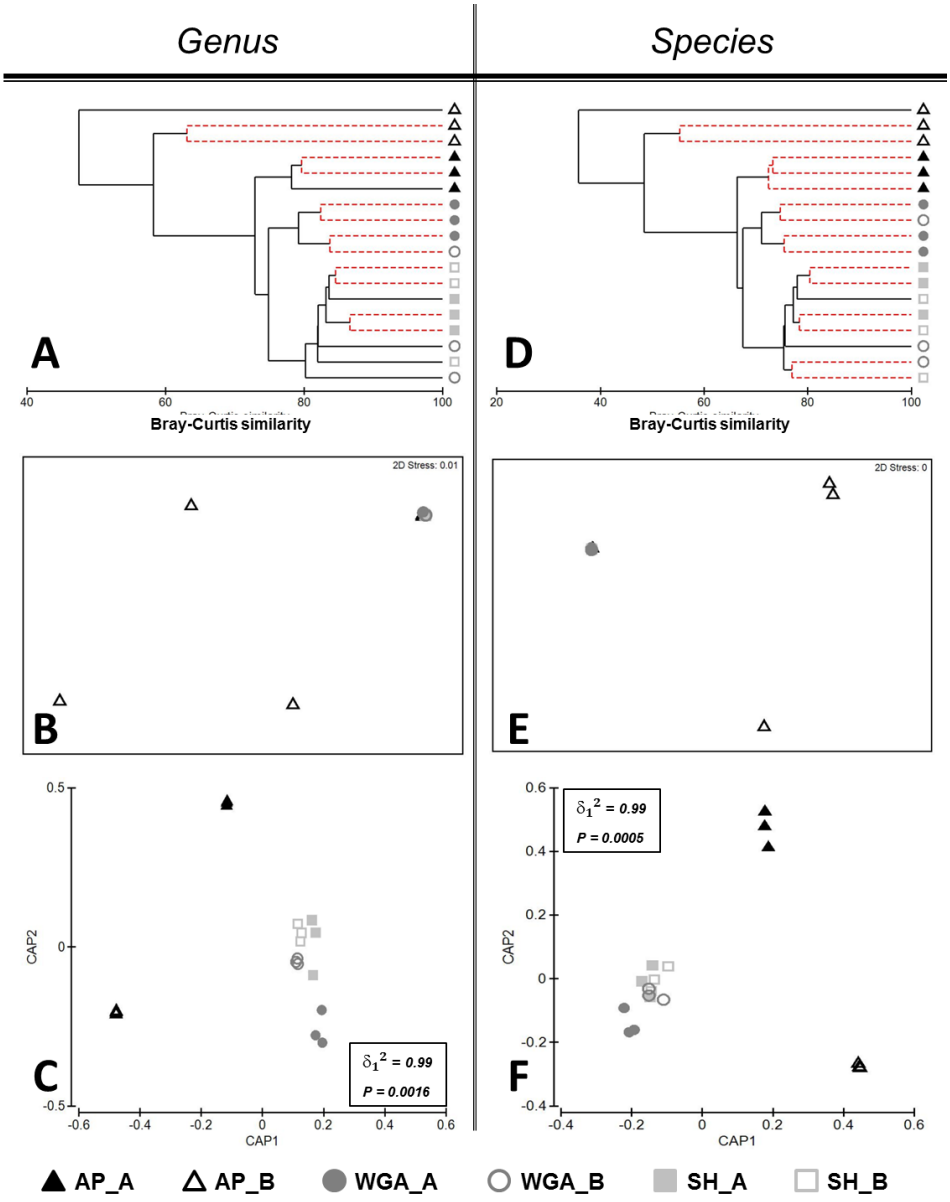



**Figure 5.17. Comparison of the taxonomic soil profiles generated from the assembled contigs at the genus (A, B, C) and species (D, E, F) resolution levels.** Bray-Curtis distance similarity matrix was calculated from the square-root transformed abundance of DNA fragments matching taxa in the M5NR database (E-value < $1\times10^{-5}$). The Bray-Curtis matrix was used for generating CLUSTER dendrogram, NMDS and CAP ordination plots. **CLUSTER analysis (A and D).** Red dashed branches on the CLUSTER dendrogram indicate no significant difference between metagenomic profiles (supported by the SIMPROF analysis, $p<0.05$). **NMDS unconstrained ordination (B and E).** The NMDS plot displays distances between samples. Data points that are closer to each other represent samples with highly similar metagenomic profiles. **CAP constrained ordination (C and F).** CAP analysis tests for differences among the groups in multivariate space. The significance of group separation along the canonical axis is indicated by the value of the squared canonical correlation ($\delta_1^2$) and P-value. A contour line on the NMDS and CAP ordinations drawn round each of the cluster defines the superimposition of clusters from CLUSTER dendrogram at the selected level of similarity.

**Table 5.17. Results of CAP model cross-validation of soil protein-derived taxonomic profiles discrimination generated from assembled contigs.**

| Original Group | AP_A | AP_B | WGA_A | WGA_B | SH_A | SH_B |
|---|---|---|---|---|---|---|
| Taxonomy level | ***Phylum*** ($m$ = 2, $\delta_1^2$ = 0.935, P = 0.0002) | | | | | |
| % correct | 100 | 67 | 100 | 0 | 67 | 33 |
| correct/total | 3/3 | 2/3 | 3/3 | 0/3 | 2/3 | 1/3 |
| Misclassified to group | n/a | SH_A | n/a | SH_A<br>SH_B<br>WGA_A | SH_B | SH_A<br>WGA_B |
| Taxonomy level | ***Class*** ($m$ = 6, $\delta_1^2$ = 0.99, P = 0.0001) | | | | | |
| % correct | 67 | 67 | 67 | 33 | 67 | 67 |
| correct/total | 2/3 | 2/3 | 2/3 | 1/3 | 2/3 | 2/3 |
| Misclassified to group | WGA_A | AP_A | SH_A | SH_A<br>SH_B | WGA_B | SH_A |
| Taxonomy level | ***Order*** ($m$ = 11, $\delta_1^2$ = 0.99, P = 0.0102) | | | | | |
| % correct | 100 | 67 | 100 | 0 | 33 | 100 |
| correct/total | 3/3 | 2/3 | 3/3 | 0/3 | 1/3 | 3/3 |
| Misclassified to group | n/a | AP_A | n/a | SH_B | SH_B<br>WGA_B | n/a |
| Taxonomy level | ***Family*** ($m$ = 5, $\delta_1^2$ = 0.99, P = 0.0001) | | | | | |
| % correct | 100 | 100 | 100 | 0 | 67 | 0 |
| correct/total | 3/3 | 3/3 | 3/3 | 0/3 | 2/3 | 0/3 |
| Misclassified to group | n/a | n/a | n/a | SH_A<br>SH_B<br>WGA_B | WGA_B | SH_A<br>WGA_B<br>WGA_B |
| Taxonomy level | ***genus*** ($m$ = 10, $\delta_1^2$ = 0.99, P = 0.0016) | | | | | |
| % correct | 100 | 100 | 67 | 0 | 67 | 33 |
| correct/total | 3/3 | 3/3 | 2/3 | 0/3 | 2/3 | 1/3 |
| Misclassified to group | n/a | n/a | WGA_B | SH_A<br>SH_B<br>WGA_A | WGA_B | WGA_A |
| Taxonomy level | ***species*** ($m$ = 6, $\delta_1^2$ = 0.99, P = 0.0005) | | | | | |
| % correct | 100 | 100 | 67 | 0 | 67 | 33 |
| correct/total | 3/3 | 3/3 | 2/3 | 0/3 | 2/3 | 1/3 |
| Misclassified to group | n/a | n/n | SH_A | WGA_A<br>SH_B<br>SH_B | WGA_B | SH_A |

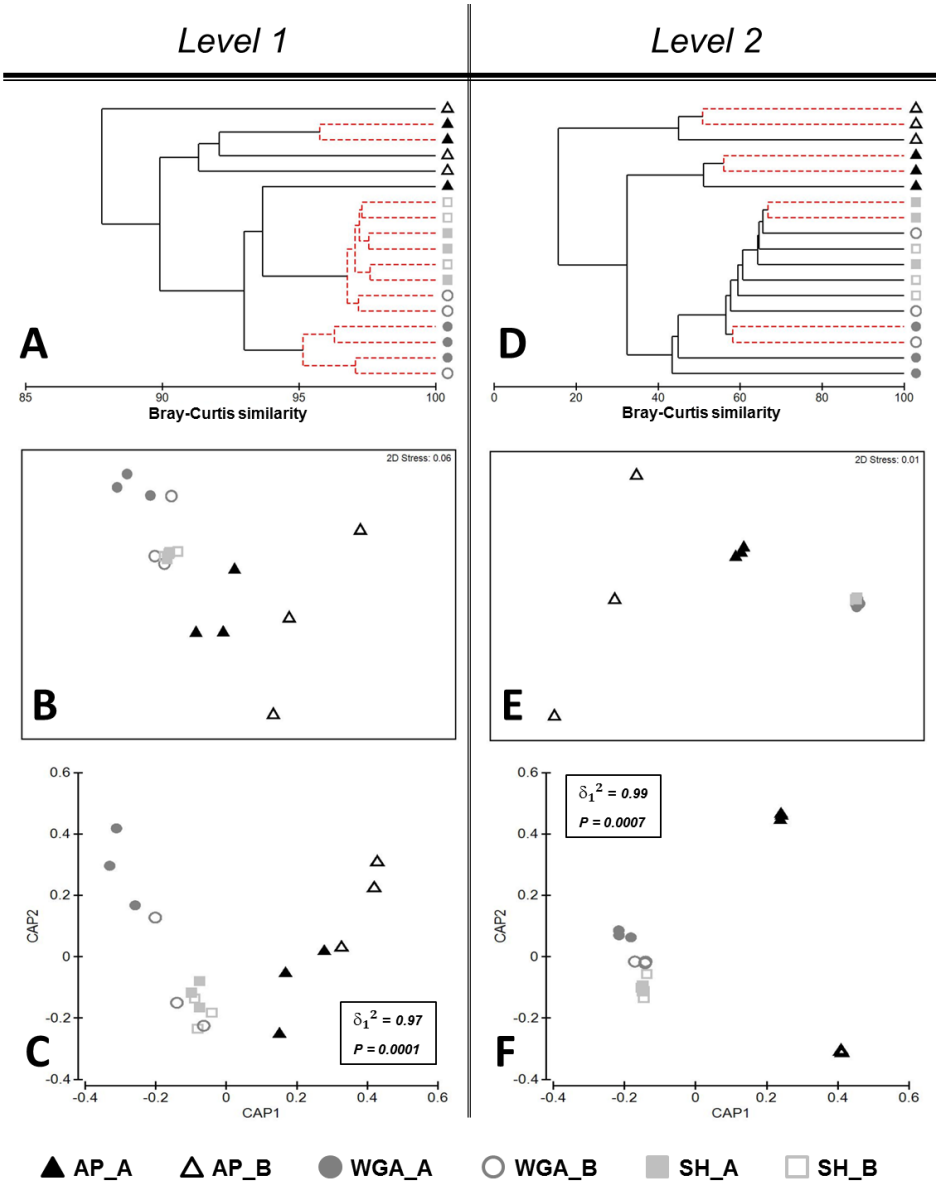



**Figure 5.18. Comparison of the metabolic soil profiles generated from the assembled contigs at the subsystems level 1 (A, B, C) and level 2 (D, E, F) resolution levels.** Bray-Curtis distance similarity matrix was calculated from the square-root transformed abundance of DNA fragments matching taxa in the SEED database (E-value < $1\times10^{-5}$). The Bray-Curtis matrix was used for generating CLUSTER dendrogram, NMDS and CAP ordination plots. **CLUSTER analysis (A and D).** Red dashed branches on the CLUSTER dendrogram indicate no significant difference between metagenomic profiles (supported by the SIMPROF analysis, $p<0.05$). **NMDS unconstrained ordination (B and E).** The NMDS plot displays distances between samples. Data points that are closer to each other represent samples with highly similar metagenomic profiles. **CAP constrained ordination (C and F).** CAP analysis tests for differences among the groups in multivariate space. The significance of group separation along the canonical axis is indicated by the value of the squared canonical correlation ($\delta_1^2$) and P-value. A contour line on the NMDS and CAP ordinations drawn round each of the cluster defines the superimposition of clusters from CLUSTER dendrogram at the selected level of similarity.

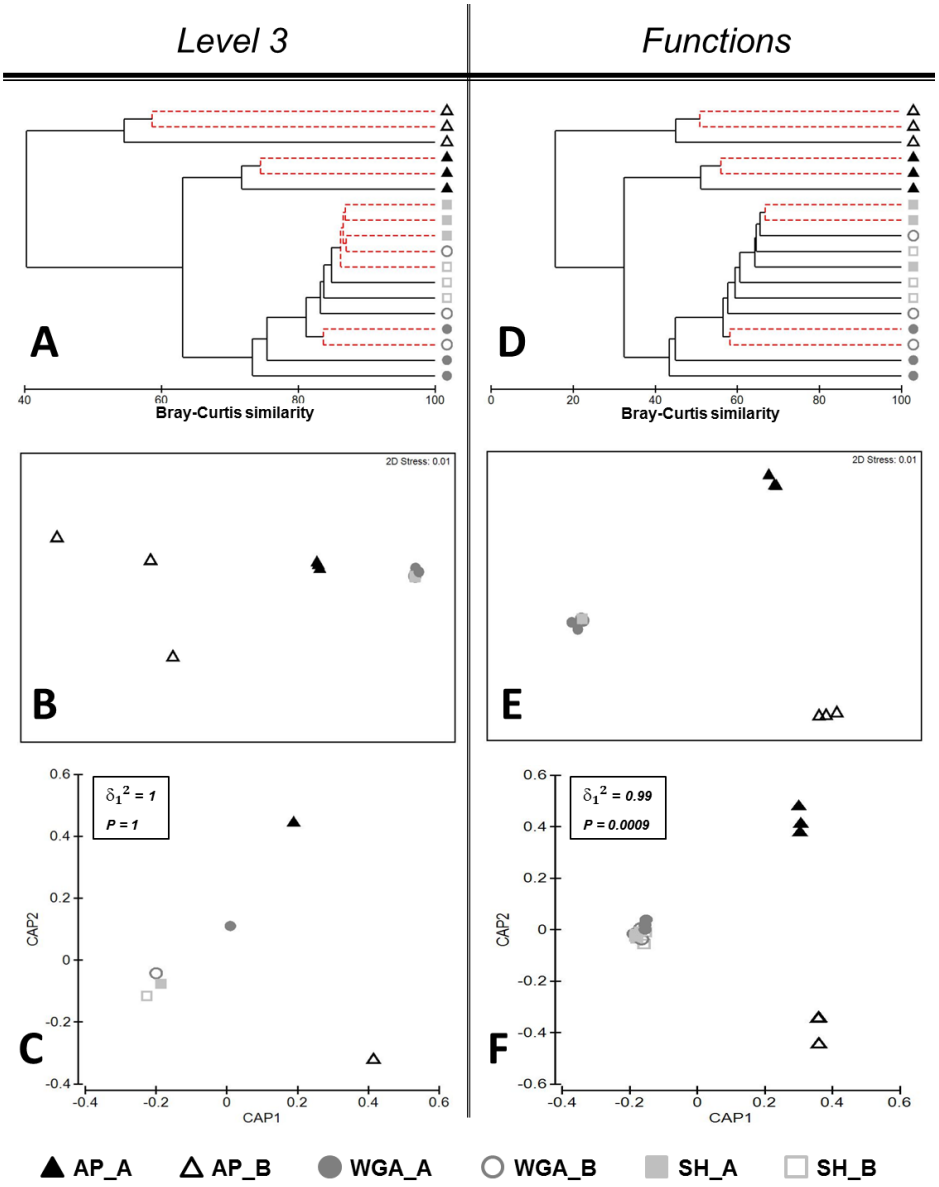



**Figure 5.19. Comparison of the metabolic soil profiles generated from the assembled contigs at the subsystems level 3 (A, B, C) and function (D, E, F) resolution levels.** Bray-Curtis distance similarity matrix was calculated from the square-root transformed abundance of DNA fragments matching taxa in the SEED database (E-value < $1\times10^{-5}$). The Bray-Curtis matrix was used for generating CLUSTER dendrogram, NMDS and CAP ordination plots. **CLUSTER analysis (A and D).** Red dashed branches on the CLUSTER dendrogram indicate no significant difference between metagenomic profiles (supported by the SIMPROF analysis, $p<0.05$). **NMDS unconstrained ordination (B and E).** The NMDS plot displays distances between samples. Data points that are closer to each other represent samples with highly similar metagenomic profiles. **CAP constrained ordination (C and F).** CAP analysis tests for differences among the groups in multivariate space. The significance of group separation along the canonical axis is indicated by the value of the squared canonical correlation ($\delta_1^2$) and P-value. A contour line on the NMDS and CAP ordinations drawn round each of the cluster defines the superimposition of clusters from CLUSTER dendrogram at the selected level of similarity.

**Table 5.18. Results of CAP model cross-validation of soil metabolic profiles discrimination generated from assembled contigs.**

| Original Group | **AP_A** | **AP_B** | **WGA_A** | **WGA_B** | **SH_A** | **SH_B** |
|---|---|---|---|---|---|---|
| Metabolic level | *level 1* ($m$ = 3, $\delta_1^2$= 0.97, P = 0.0001) | | | | | |
| % correct | 100 | 33 | 100 | 0 | 33 | 67 |
| correct/total | 3/3 | 1/3 | 3/3 | 0/3 | 1/3 | 2/3 |
| Misclassified to group | n/a | AP_A | n/a | SH_A<br>SH_G<br>WGA_A | SH_B<br>WGA_B | SH_A |
| Metabolic level | *level 2* ($m$ = 8, $\delta_1^2$ = 0.99, P = 0.0007) | | | | | |
| % correct | 100 | 100 | 67 | 0 | 100 | 67 |
| correct/total | 3/3 | 3/3 | 2/3 | 0/3 | 3/3 | 2/3 |
| Misclassified to group | n/a | n/a | WGA_B | WGA_A<br>WGA_A<br>SH_B | n/a | WGA_A |
| Metabolic level | *level 3* ($m$ = 14, $\delta_1^2$ = 1, P = 1) | | | | | |
| % correct | 100 | 100 | 67 | 33 | 100 | 0 |
| correct/total | 3/3 | 3/3 | 2/3 | 1/3 | 3/3 | 0/3 |
| Misclassified to group | n/a | n/a | WGA_B | WGA_A<br>SH_A | n/a | SH_A |
| Metabolic level | *functions* ($m$ = 8, $\delta_1^2$ = 0.99, P = 0.0009) | | | | | |
| % correct | 100 | 100 | 67 | 0 | 100 | 33 |
| correct/total | 3/3 | 3/3 | 2/3 | 0/3 | 3/3 | 1/3 |
| Misclassified to group | n/a | n/a | WGA_B | SH_A | n/a | SH_A |

**Comparison of taxonomic profiles based on rRNA gene fragment classification.**

CLUSTER analysis and NMDS ordination of rRNA-based taxonomic profiles at all levels of taxonomy demonstrated the formation of a heterogeneous mixed cluster of the profiles from the samples included in the SH- and WGA-based groups (Figure 5.20 - Figure 5.22). CAP analysis statistics of group discrimination showed insignificance of the obtained results with P-value more than 5% (Figure 5.20 - Figure 5.22).

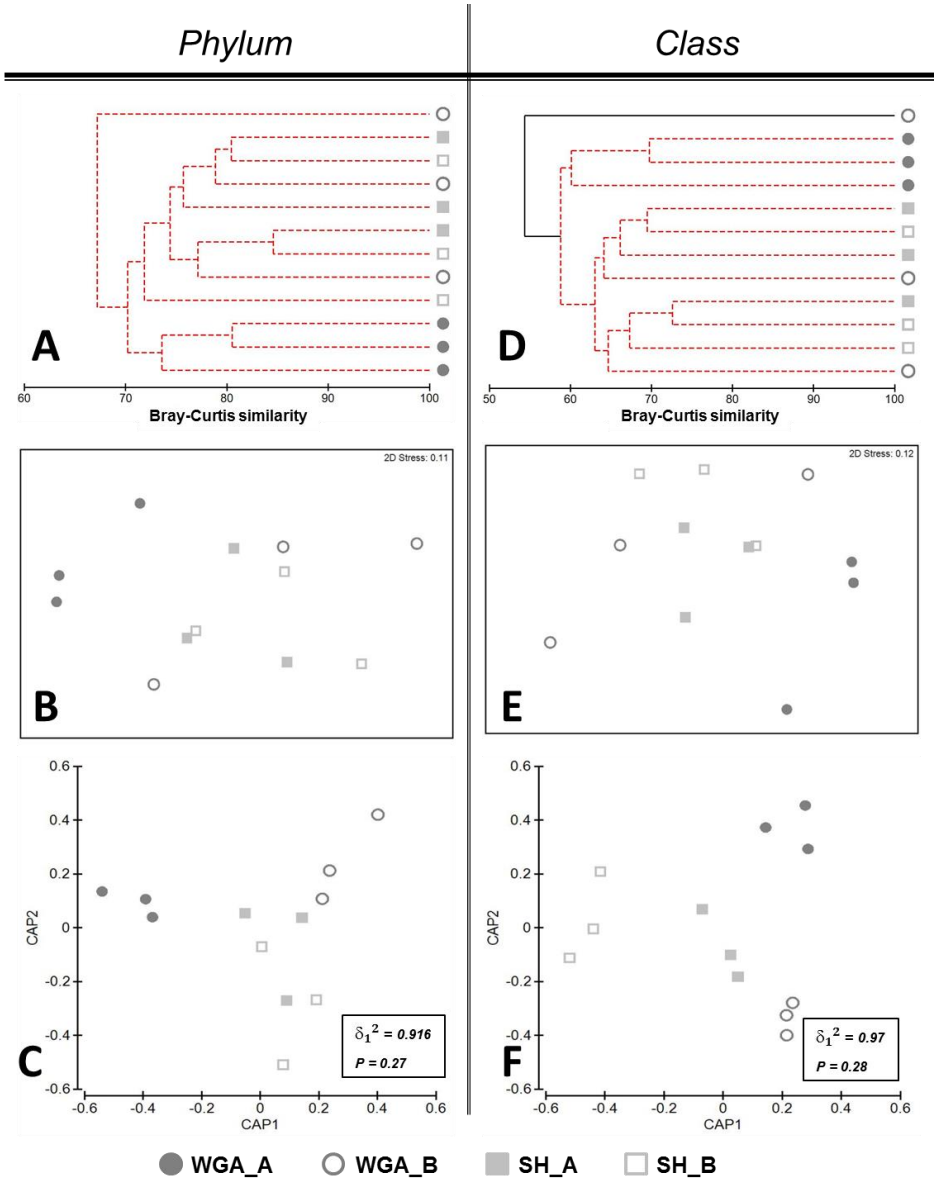



**Figure 5.20. Comparison of the soil rRNA profiles generated on the assembled contigs at the phylum (A, B, C) and class (D, E, F) taxonomic resolution levels.** Bray-Curtis distance similarity matrix was calculated from the square-root transformed abundance of DNA fragments matching taxa in the M5RNA database (E-value < $1\times10^{-5}$). The Bray-Curtis matrix was used for generating CLUSTER dendrogram, NMDS and CAP ordination plots. **CLUSTER analysis (A and D)**. Red dashed branches on the CLUSTER dendrogram indicate no significant difference between metagenomic profiles (supported by the SIMPROF analysis, $p<0.05$). **NMDS unconstrained ordination (B and E)**. The NMDS plot displays distances between samples. Data points that are closer to each other represent samples with highly similar metagenomic profiles. **CAP constrained ordination (C and F)**. CAP analysis tests for differences among the groups in multivariate space. The significance of group separation along the canonical axis is indicated by the value of the squared canonical correlation ($\delta_1^2$) and P-value ($P<0.05$). A contour line on the NMDS and CAP ordinations drawn round each of the cluster defines the superimposition of clusters from CLUSTER dendrogram at the selected level of similarity.

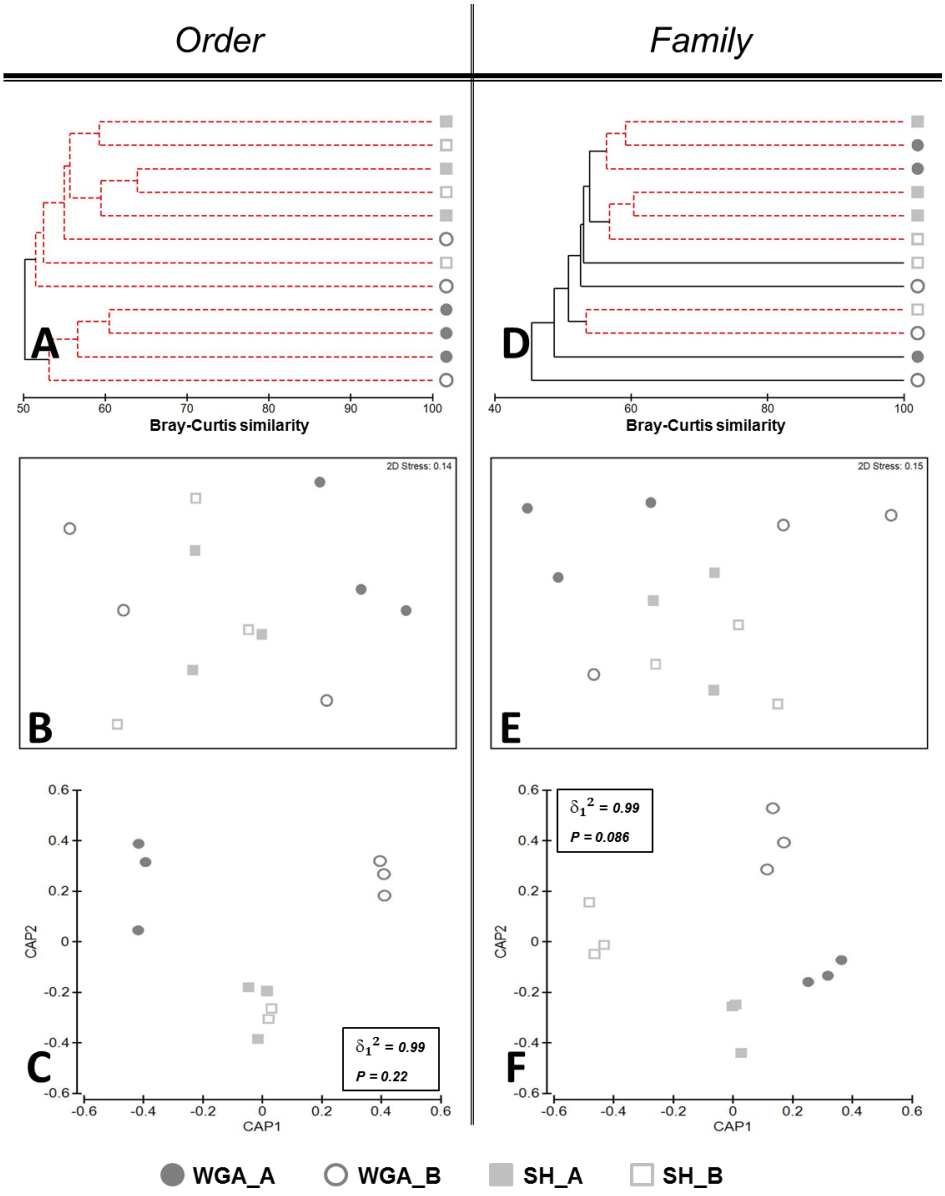



**Figure 5.21. Comparison of the soil rRNA profiles generated on the assembled contigs at the order (A, B, C) and family (D, E, F) taxonomic resolution levels.** Bray-Curtis distance similarity matrix was calculated from the square-root transformed abundance of DNA fragments matching taxa in the M5RNA database (E-value < $1\times10^{-5}$). The Bray-Curtis matrix was used for generating CLUSTER dendrogram, NMDS and CAP ordination plots. **CLUSTER analysis (A and D)**. Red dashed branches on the CLUSTER dendrogram indicate no significant difference between metagenomic profiles (supported by the SIMPROF analysis, $p<0.05$). **NMDS unconstrained ordination (B and E)**. The NMDS plot displays distances between samples. Data points that are closer to each other represent samples with highly similar metagenomic profiles. **CAP constrained ordination (C and F)**. CAP analysis tests for differences among the groups in multivariate space. The significance of group separation along the canonical axis is indicated by the value of the squared canonical correlation ($\delta_1^2$) and P-value ($P<0.05$). A contour line on the NMDS and CAP ordinations drawn round each of the cluster defines the superimposition of clusters from CLUSTER dendrogram at the selected level of similarity.

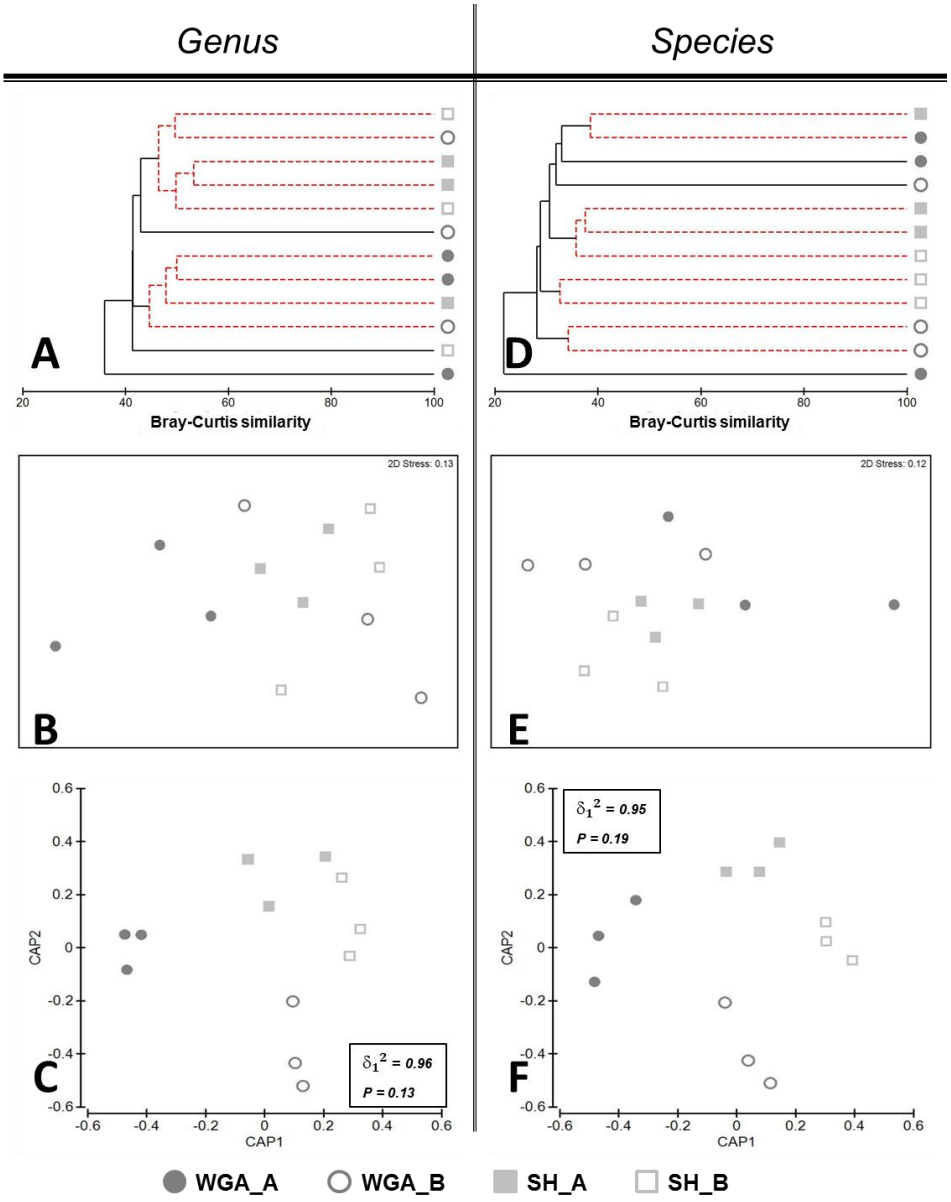



**Figure 5.22. Comparison of the soil rRNA profiles generated on the assembled contigs at the genus (A, B, C) and species (D, E, F) taxonomic resolution levels.** Bray-Curtis distance similarity matrix was calculated from the square-root transformed abundance of DNA fragments matching taxa in the M5RNA database (E-value < $1\times10^{-5}$). The Bray-Curtis matrix was used for generating CLUSTER dendrogram, NMDS and CAP ordination plots. **CLUSTER analysis (A and D)**. Red dashed branches on the CLUSTER dendrogram indicate no significant difference between metagenomic profiles (supported by the SIMPROF analysis, $p<0.05$). **NMDS unconstrained ordination (B and E)**. The NMDS plot displays distances between samples. Data points that are closer to each other represent samples with highly similar metagenomic profiles. **CAP constrained ordination (C and F)**. CAP analysis tests for differences among the groups in multivariate space. The significance of group separation along the canonical axis is indicated by the value of the squared canonical correlation ($\delta_1^2$) and P-value ($P<0.05$). A contour line on the NMDS and CAP ordinations drawn round each of the cluster defines the superimposition of clusters from CLUSTER dendrogram at the selected level of similarity.

Despite the low species coverage derived from rarefaction analysis of soil metagenomic profiles produced by the AP-based approach it resulted in a 100% correct discrimination between soil samples from different locations. In particular, AP-PCR-based whole metagenome sequencing approach was able to discriminate visually similar soil samples based on differences in both taxonomic and metabolic compositions at all levels of classification. This might be explained by unexplored pre-enrichment mechanism of AP-PCR that is based on the primer sequence targeting both dominant and rare microorganisms.

SH- and WGA-based metagenomic sequencing approaches showed incorrect and inconsistent separation of soil samples according to their sampling sites using taxonomic (protein and ribosomal) and metabolic classifications. This was shown both for the individual reads and assembled contigs. Comparison of the SH- and WGA-based profiles revealed not only misclassification of the samples between the locations but often between repeat analyses of each sequencing approach. The profiles from WGA_A group were the only exception, as the profiles had a 100% correct allocation success under cross validation of the CAP model. The similarity of the profiles generated by these methods appears to be driven by the highly similar, or even identical, dominant microorganisms found in the soil samples collected from two distinct sites of similar urban type. This supports the theory that the data generated by shotgun sequencing are commonly shifted towards describing the most abundant taxa leaving the contribution of rare microorganisms undervalued for comparative analysis (Zarraonaindia et al. 2013).

## 5.4 Conclusion

The research presented in this chapter focuses on assessing the ability of SH and WGA random whole metagenomic sequencing approaches, which are widely accepted as the most comprehensive sources of data for studying complex microbial communities (Howe, Jansson, S. A. Malfatti, et al. 2014), to describe and differentiate soils from two similar parklands approximately 3 km apart within Adelaide residential area. The vegetation categories of the chosen locations appeared to be very similar, with widespread grass and trees species. The SH and WGA sequencing approaches were consequently compared with AP-PCR-based sequencing that had been presented in Chapter 4.

The composition of the soil microbial communities was determined from both taxonomic classification of rRNA fragments and the taxonomic and metabolic assignment of functional gene fragments for the initial and assembled datasets. Similar taxonomic distribution of dominant Phyla was observed across all metagenomic datasets, including assembled ones using these two different annotation pipelines.

The comparison of metagenomic profiles was performed with a number of unconstrained statistical tools including CLUSTER and NMDS analyses. Additionally constrained CAP analysis, previously shown to be successful at soil microbial communities discrimination, was used as it challenges a predefined hypothesis of ‘no difference’ between groups of samples separated by factors of sampling sites and sequencing method applied (Aderson & Willis 2003; Smith et al. 2013).

The shotgun and WGA-based approaches generated highly similar metagenomic profiles for soil samples such that the soil samples could not be distinguished. An AP-PCR-based approach was shown to be the most powerful technique for obtaining site-

specific metagenomic DNA profiles which were able to successfully discriminate between similar soil samples taken from different locations.

# Chapter 6. Reference-independent comparative metagenomics for forensic soil analysis

## 6.1 Introduction

Current analysis of metagenomic sequencing datasets is commonly achieved by their comparison to appropriate reference databases. These include protein databases which are used mostly in the analysis of random whole metagenomic sequencing data (such as SEED (Overbeek et al. 2005), M5NR (Meyer et al. 2008)) and ribosomal databases used for gene-targeted sequencing (Greengenes (http:// greengenes.lbl.gov) (DeSantis et al. 2006), RDP (http://rdp.cme.msu.edu/) (Cole et al. 2014) and SILVA (www.arb-silva. de/) (Quast et al. 2013)). The output is the generation of taxonomic and metabolic profiles. Comparison of the taxonomic or functional metagenomic profiles of two soil samples aims to give a measure of the level of relatedness between the samples. However these reference databases represent only a small fraction of the biological diversity of such a complex microbial system as soil. This is because the majority of reference genomes included in the databases are derived from known/cultivated microorganisms. This is supported by the proposition that more than 99% of all microbial species from the soil environment are 'uncultivable' (Torsvik & Øvreås 2002). Thus, metagenomic reads associated with known microorganisms can be easily identified and analysed using fast homology search algorithms and a suitable reference database. Due to the incomplete nature of current databases, a significant part of reads, which might bear a valuable piece of information, are therefore left unannotated and often disregarded in further comparative analyses resulting in biased conclusions (Carlos et al. 2014). Often *de novo* assembly of all sequencing reads helps to reduce the complexity of the metagenomic datasets and to link reads of unknown origin to those that can be annotated (Howe et al. 2014).

The problem of numerous unannotated reads that remain after taxonomic or metabolic classification is highlighted by the observation in the current study that only

25-35% of total features found were assigned to the MG-RAST M5NR database (see Chapters 4 & 5). This fits well with published data showing that only a small fraction of sequencing reads, about 20-40%, can be mapped to database refereed genes and proteins (Gilbert et al. 2010).

One feasible way to compare whole metagenomic sequence datasets is therefore to identify reads that are shared or similar between samples without requiring a database of reference sequences. Such comparative analysis of sequence datasets is named reference-independent comparative metagenomics or sequence signature-based comparison (Jiang et al. 2012). These types of methods can potentially use all sequencing reads available and therefore can theoretically reveal more accurate relationships between samples.

Maillet *et al.* has recently published an algorithm for *de novo* pairwise comparison of metagenomic datasets (Compareads) (Maillet et al. 2012). The rationale of the method is based on the computation of similarity between two raw sequencing datasets. According to the proposed approach two reads are considered as similar if they share *t* non-overlapping *k*-mers, where the *k*-mer is a DNA sequence of *k* nucleotides long, and *t* is the number of *k*-mers to be found in the reads being compared. An example of this approach was that of Tas *et al.* in the investigation of the microbial community structure of permafrost soils affected by fire (Taş et al. 2014). Without prior knowledge of taxonomic composition of the soil layers, the authors showed drastic changes in microbial composition of the surface layer of the soil affected by fire (within layer similarity of 1.67%) compared with the control site (within surface layer similarity of 25.33%).

Another reference-independent comparative metagenomics approach based on cross-assembly (CRASS) was proposed by Dutilh (Dutilh et al. 2012). The method

involves building a common cross-assembly from all the initial datasets and then the distances between the metagenomes are calculated based on scoring the number of reads from each metagenome that were included in the final cross-assembly. The method has been applied successfully for comparison of microbiomes derived from six specimens of marine sponge *A. brasiliensis* with 23 other marine animal microbiomes (Trindade-Silva et al. 2012). It was found that the *A. brasiliensis* had a species-specific microbiome compared to those of the other marine animals including the Australian sponge *Cymbastela concentrica*, healthy and morbid fish, the mussel species *Mytilus galloprovincialis* and *M. eduli*. This result has also been confirmed by annotation of sequences with GenBank and SEED Subsystems databases using the MG-RAST on-line server. Similar work showed feasibility of the CRASS method for comparison of viral biomes (viromes) from six different *Hydra* species (Grasis et al. 2014). The authors concluded that each of six *Hyrda* species studied had a unique viral community structure.

The goal of this chapter is to assess the performance of reference-independent sequence dataset comparison algorithms for discrimination between visually similar soils collected from two distinct urban sites of similar land use only 3 km apart. Metagenomic datasets obtained by SH-, WGA-, and AP-based sequencing were taken from Chapter 5, where they were compared using a traditional reference database annotation approach and where correct allocation of the soil samples according to their sampling sites was achieved only using AP-based generated metagenomic profiles. This chapter evaluates and compares two reference-independent algorithms: CRASS (Dutilh et al. 2012) and Compareads (Maillet et al. 2012).

## 6.2 Materials and methods

### 6.2.1 Sequence datasets

In this chapter, raw (unprocessed) sequencing datasets generated on the soil DNA samples Aw2 – Aw4 and Bw2 – Bw4 (Table 4.2, Chapter 4) by SH-, WGA- and AP-PCR based sequencing techniques (for details see Chapter 5) were used. In total 18 sequencing datasets (2 locations A & B × 3 replicates × 3 sequencing approaches) were subjected to analyses.

### 6.2.2 Quality filtering of sequencing data, primers trimming and sub-sampling.

Cutadapt v1.1 tool (Martin 2011) was used to trim AP-PCR primer sequence from the raw sequencing reads of the AP-based dataset using a strict zero mismatch threshold (parameters; -b [Forward; CCCTCGAACACCACCTCC] –b [Reverse complement; GGAGGTGGTGTTCGAGGG] –m [50 # discard reads that are shorter than min-length].

Fastx_clipper tool (FASTX-toolkit v 0.0.14; http://hannonlab.cshl.edu/fastx toolkit) was subsequently used to remove reads less than 50 bp in length (parameters; -l 50) from SH- and WGA-based datasets.

A Fastq_quality_filter tool was used to remove the reads with a Phred quality score less than 20 for 90% of the read (parameters; -q 20 –p 90).

Python script Subsampler.py was used across the three methods tested to rarefy the data to ensure an even sequencing depth for each sample for the subsequent comparative analysis

(https://github.com/macmanes/error_correction/blob/master/scripts/subsampler.py). Subsampling of sequencing reads was performed in a Linux operational system. All procedures were performed according to developer instructions. The command to execute the script was: 'python subsampler.py Input_file_name.fasta N > Output_file_name'; where N is a desirable number of reads in the output file.

### 6.2.3 Reference-independent analysis of sequencing data

**Cross-Assembly of Metagenomes (CRASS).**

All initial metagenomic datasets (quality filtered and subsampled) to be analysed were combined in one single file containing all reads using Geneious R7 software (BioMatters, New Zealand). Then the obtained file was subjected to an assembly procedure using gsAssembler (Roche Corp., Switzerland) with default settings and a single ACE file as an output file format. The obtained ACE file along with initial datasets were then uploaded to the CRASS web-site (http://edwards.sdsu.edu/crass/) and the CRASS routine was launched. Distance matrices of all pairwise comparisons of metagenomic datasets were calculated using SHOT and READS formulae and then were visualised as cladograms from the CRASS output.

**A *de novo* comparative metagenomics analysis algorithm (Compareads).**

Compareads was used to compute the pairwise similarity scores between metagenomic datasets. The program was run in a command line in the Linux operational system using commands provided by the developers (Maillet et al. 2012). Metagenomic comparison was performed with default Compareads parameters ($k$ = 33, $t$ = 2) as well as with varying parameters of $k$ value from 15 to 33 and $t$-value 1 and 2.

### 6.2.4 Statistical analysis

CLUSTER and NMDS tools from Primer 6 statistical package were used for the analysis of Compareads results.

## 6.3 Results and discussion

Primer trimming and quality filtering of raw sequencing datasets were performed with the open source software described in the Materials and Methods section. For each soil DNA sample three datasets were generated from the same DNA template using three sequencing approaches, namely shotgun (SH), whole genome amplification (WGA) and arbitrary primed PCR (AP-PCR). SH sequencing resulted in an average of 672,542 (531,108 – 806,843) sequence reads with an average sequence length of 198 ± 73 bases for a total of > 133 Mbp of sequence. The WGA dataset consisted of an average of 911,554 (506,028 – 2,012,359) sequences with an average of 198 ± 75 bases in length for a total of > 178 Mbp. The AP-based approach gave an average of 468,187 (74,370 – 1,047,266) reads with a mean of 143 ± 69 bases in length for a total of >70.7Mbp (Table 5.2), as was described in Chapter 5.

Approximately 6% of reads on average were eliminated from AP-based datasets during the primer trimming procedure and 45% of low-quality reads were then removed at the filtering step (Table 6.1). Quality filtering of SH and WGA based datasets resulted in the exclusion of 73% of reads. A quality filtering procedure insured that no reads containing more than 10% of nucleotides with Phred score less than 20, indicating an accuracy of 99% base calling, were included in the consecutive analysis.

**Table 6.1. General characteristics of AP-, SH- and WGA-based sequence datasets.**

| Sequencing approach | Average number of reads (range) | Number of Mbp | Average read length, bp ± SD | Trimmed primers, % | Eleminated reads with length < 50 nt, % | Quality filtered reads (Q20), % | Number of reads left after QF (range) |
|---|---|---|---|---|---|---|---|
| SH | 672,542 (531,108 – 806,483) | 133.6 | 198 ± 73 | N/A | 16 | 57 | 186,365 (112,979-2,30,967) |
| AP | 468187 (74,370 – 1,074,266) | 70.7 | 142 ± 69 | 6 | N/A | 45 | 203,345 (42,114 – 400,246) |
| WGA | 911553 (506,028 – 2,012,359) | 178.5 | 198 ± 75 | N/A | 16 | 57 | 230,276 (153,348 – 376,685) |

It has been noted that the Compareads approach is sensitive to the number of reads used in the analysis, where a considerable difference in the number of sequences per sample results in false similarity estimates (Taş et al. 2014). Because of this the comparison was carried out on randomly subsampled datasets at two sequencing depths of 40000 and 4000 reads.

### 6.3.1 Comparison of metagenomic datasets using Compareads algorithm

The Compareads algorithm is an intuitively straight forward process of comparing raw sequencing datasets. It relies on finding and calculating the number of similar reads between two datasets being compared (Maillet et al. 2012). According to the Compareads default parameters, two reads were assumed to be similar if they shared at least two *k*-mers of 33 nucleotides length. Each comparison of two sequencing datasets by Compareads resulted in a pairwise similarity score. Results of the pairwise comparisons between all datasets with a sequencing depth of 40,000 reads using the default parameters ($k$ = 33 and $t$ = 2) were then summarised as a similarity matrix (Table 6.2) to be used for subsequent hierarchical agglomerative clustering (CLUSTER) and non-metric multidimensional scaling (NMDS).

**Table 6.2. Pairwise similarity scores between metagenomic datasets obtained using Compareads software at default parameters (*k* = 33, *t* = 2).**

| | SH_ Aw2 | SH_ Aw3 | SH_ Aw4 | SH_ Bw2 | SH_ Bw3 | SH_ Bw4 | AP_ Aw2 | AP_ Aw3 | AP_ Aw4 | AP_ Bw2 | AP_ Bw3 | AP_ Bw4 | WGA_ Aw2 | WGA_ Aw3 | WGA_ Aw4 | WGA_ Bw2 | WGA_ Bw3 | WGA_ Bw4 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **SH_Aw2** | | | | | | | | | | | | | | | | | | |
| **SH_Aw3** | 0.2 | | | | | | | | | | | | | | | | | |
| **SH_Aw4** | 0.3 | 0.3 | | | | | | | | | | | | | | | | |
| **SH_Bw2** | 0.2 | 0.3 | 0.2 | | | | | | | | | | | | | | | |
| **SH_Bw3** | 0.1 | 0.2 | 0.1 | 0.2 | | | | | | | | | | | | | | |
| **SH_Bw4** | 0.2 | 0.2 | 0.2 | 0.3 | 0.2 | | | | | | | | | | | | | |
| **AP_Aw2** | 0.0 | 0.0 | 0.5 | 0.0 | 0.0 | 0.0 | | | | | | | | | | | | |
| **AP_Aw3** | 0.0 | 0.0 | 0.9 | 0.0 | 0.0 | 0.0 | 65.1 | | | | | | | | | | | |
| **AP_Aw4** | 0.1 | 0.0 | 1.4 | 0.0 | 0.0 | 0.0 | 56.6 | 70.1 | | | | | | | | | | |
| **AP_Bw2** | 0.0 | 0.0 | 0.0 | 0.1 | 0.0 | 0.0 | 51.8 | 56.9 | 33.1 | | | | | | | | | |
| **AP_Bw3** | 0.0 | 0.0 | 0.0 | 0.1 | 0.0 | 0.0 | 46.6 | 55.2 | 29.1 | 82.5 | | | | | | | | |
| **AP_Bw4** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 51.2 | 55.2 | 31.8 | 81.6 | 80.3 | | | | | | | |
| **WGA_Aw2** | 0.1 | 0.2 | 0.2 | 0.1 | 0.1 | 0.1 | 0.0 | 0.1 | 0.1 | 0.0 | 0.0 | 0.0 | | | | | | |
| **WGA_Aw3** | 0.2 | 0.2 | 0.3 | 0.1 | 0.1 | 0.1 | 0.1 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 11.9 | | | | | |
| **WGA_Aw4** | 0.2 | 0.1 | 0.3 | 0.1 | 0.1 | 0.1 | 0.0 | 0.1 | 0.2 | 0.0 | 0.0 | 0.0 | 9.8 | 12.4 | | | | |
| **WGA_Bw2** | 0.2 | 0.2 | 0.2 | 0.4 | 0.2 | 0.3 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.1 | 0.1 | 0.1 | | | |
| **WGA_Bw3** | 0.1 | 0.1 | 0.1 | 0.2 | 0.2 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 3.9 | 1.7 | 0.7 | 0.6 | | |
| **WGA_Bw4** | 0.1 | 0.2 | 0.1 | 0.2 | 0.2 | 0.2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.6 | 0.4 | 0.1 | 0.6 | 1.5 | |

The AP-based datasets formed two separate clusters according to their sampling locations as clearly seen on the dendrogram (Figure 6.1A) and NMDS ordination plot (Figure 6.1B) with an average similarity of 63.9 ± 6.8% for the AP_A dataset and 81.5 ± 1.1% for the AP-B datasets (Table 6.3). WGA-based datasets from location A had an average similarity of only 11.4 ± 1.4% as opposed to a similarity of 0.9 ± 0.5% for the WGA_B dataset; this allowed for a formation of a separate cluster as seen on the dendrogram (Figure 6.1A). By contrast, SH-based datasets showed similar low similarity scores, <0.3%, for both within site and between sites comparisons. The result was that separate clusters did not formed and did not allow for the discrimination of soils samples taken from locations A and B.

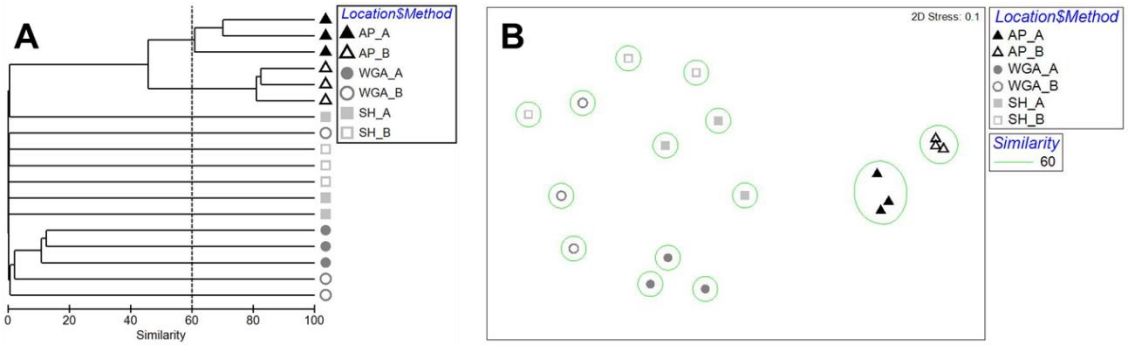



**Figure 6.1. Comparison of the metagenomic datasets using multivariate statistical tools: (A) hierarchical agglomerative clustering (CLUSTER) and (B) non-metric multidimensional scaling (NMDS).** Pairwise similarity matrix (Table 6.2) obtained using Compareads with default parameters was used for generating CLUSTER dendrogram and NMDS ordination plot. A contour line on the NMDS plot drawn round each of the clusters defines the superimposition of clusters from CLUSTER dendrogram at the selected level of similarity.

**Table 6.3. Average similarity scores calculated for metagenomic datasets obtained by different sequencing approaches from the same soils samples and compared by Compareads with default parameters.**

| Dataset Name | AP_Aw2-4 | AP_Bw2-4 | SH_Aw2-4 | SH_Bw2-4 | WGA_Aw2-4 | WGA_Bw2-4 |
|---|---|---|---|---|---|---|
| **AP_Aw2-4** | 63.9 ± 6.8 | | | | | |
| **AP_Bw2-4** | 45.7 ± 11.2 | 81.5 ± 11 | | | | |
| **SH_Aw2-4** | | | 0.3 ± 0.02 | | | |
| **SH_Bw2-4** | | | 0.2 ± 0.05 | 0.1 ± 0.06 | | |
| **WGA_Aw2-4** | | | | | 11.4 ± 1.4 | |
| **WGA_Bw2-4** | | | | | 0.8 ± 1.3 | 0.9 ± 0.5 |

**Comparison of datasets at varying parameters of *t* and *k* values**

It was expected that varying the *k* and *t* parameters of the Compareads algorithm should affect its ability to discriminate soil samples taken from the same or different sites. Thus pairwise similarity scores between datasets were also calculated for varying parameters of *k*-mer length (from $k = 15$ to $k = 33$) and *t*-value (1 and 2, i.e. how many times the *k*-mer has to be found). The average similarity scores obtained for within site and between sites comparisons of the metagenomic datasets generated by each of the sequencing approaches were plotted as a function of *k*-mer length for two values of $t = 1$ and $t = 2$. Figure 6.2 shows that for the parameter $t = 1$, all the curves start from the similarity score of 100% at $k = 15$ for both within and between sites comparisons for all datasets generated by the three sequencing approaches. The results indicate that at $k = 15$ none of the SH-, WGA-based, or AP-based dataset allowed for discrimination of the soils. Interestingly an increase in the *k* value led to the gradual decrease of the similarity level between the datasets but in a different way for each of the three sequencing approaches. In particular the average similarity scores between AP-based datasets from the samples collected within a site decreased down to 76% at $k = 25$ and then remained stable for all greater *k* values as indicated by the plateau region on the curve (black line, Figure 6.2A). The average similarity scores between AP-based datasets corresponding

to soils from different sites reached a similarity plateau of 52% at $k$ = 25 (red line, Figure 6.2A). The absence of an overlap of similarity scores for comparison of within and between sites across all values of $k$ indicated a reliable discrimination between visually similar soils by the AP-PCR based sequencing approach.

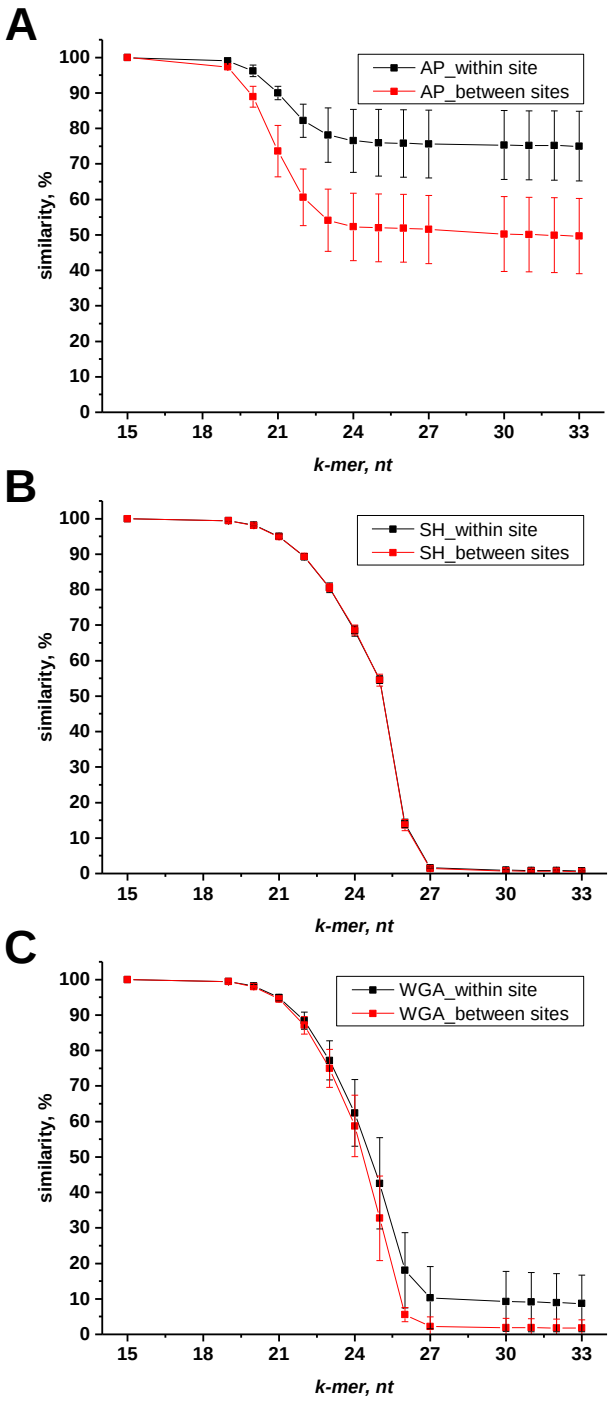



**Figure 6.2. Visualisation of the average similarity score changes obtained by the Compareads software with *t* =1 and varying *k* for the within site and between sites comparison of metagenomic datasets generated by three sequencing approaches: AP (A), SH (B) and WGA (C).**

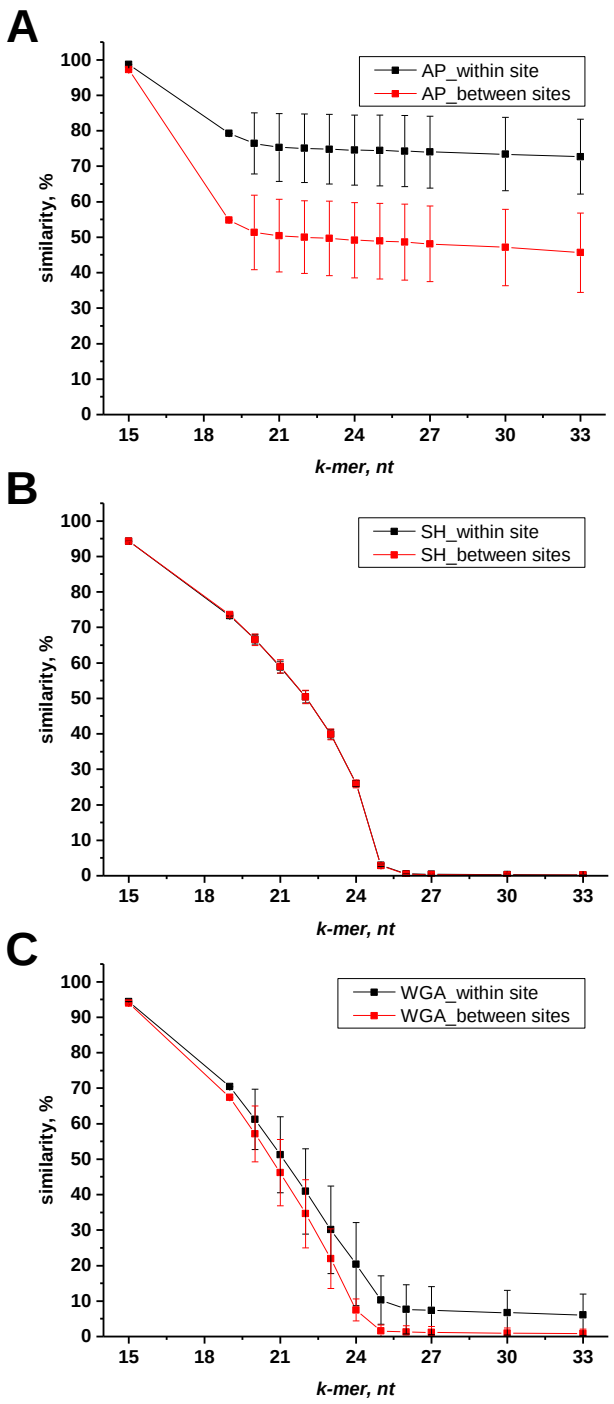



**Figure 6.3. Visualisation of the average similarity score changes obtained by the Compareads software with t =2 and varying k for the within site and between sites comparison of metagenomic datasets generated by three sequencing approaches: AP (A), SH (B) and WGA (C).**

In contrast comparison of the SH-based datasets resulted in identical similarity scores obtained for within and between sites comparison at all *k* values tested, as illustrated by the complete overlap of the similarity curves corresponding to the SH-based datasets on the plot (Figure 6.2B). It is of note that when $k = 30$ or greater the average similarity level between SH-based datasets decreased below 1%. An analogous behaviour of the similarity curves was observed for the WGA-based datasets. On average 11% of similarity between WGA_A datasets was only 2% greater than the average similarity scores for soils taken within a site (approximately 9%) but the overlapping average similarity score distributions for within site and between sites comparisons did not allow for reliable discrimination of soils (Figure 6.2C).

Figure 6.3 shows the result of the metagenomic dataset comparison at $t = 2$ and varying *k* value parameters. This figure illustrates that at a *k* value of 20 the AP-based datasets reached a plateau at 76% for the within site and 51% for the between sites comparison (Figure 6.3A). This clearly indicated that at $t = 2$ the lower *k*-mer length can be efficiently used for reliable differentiation of AP-based datasets obtained from visually similar soils. Whereas the curves corresponding to the SH- and WGA-based datasets reached plateau with similarity scores of less than 1% at $k = 26$, except for the WGA based within site comparison at 8% (Figure 6.3B and Figure 6.3C).

**Comparison of datasets at different sequencing depth.**

As evident from the previous section the Compareads algorithm was successful at discrimination of AP-based datasets using a sequencing depth of 40,000 reads. Corresponding values of average similarity scores for within site and between sites comparisons are presented in Table 6.4. In order to evaluate how the decrease in the number of reads in the analysed datasets effects the performance of the Compareads, the

comparison of the datasets randomly subsampled down to 4,000 reads was conducted using the default settings ($t$ = 2 and $k$ = 33). From Table 6.4 it can be seen that for the AP-based datasets the analysis showed efficient discrimination of the soils supported by within site and between sites average similarity scores of 61.5 ± 14.2% and 36.5 ± 10.7 %, respectively. It was not possible to reliably differentiate soils using SH- and WGA-based datasets because of the low similarity scores, <1%, obtained for both within site and between sites comparisons, except for the 3% similarity for the WGA within site comparison.

**Table 6.4. Average similarity scores calculated for metagenomic datasets at different sequencing depths corresponding to soils samples originating from a common site or different sites compared by Compareads with default parameters.**

| Number of reads | Comparison | AP | SH | WGA |
|---|---|---|---|---|
| 40,000 | Within a site | 72.7 ± 10.6 | 0.3 ± 0.04 | 6.1 ± 5.8 |
| | Between sites | 45.7 ± 11.2 | 0.2 ± 0.05 | 0.8 ± 1.3 |
| 4,000 | Within a site | 61.5 ± 14.2 | 0.05 ± 0.02 | 2.7 ± 2.9 |
| | Between sites | 36.5 ± 10.7 | 0.04 ± 0.03 | 0.2 ± 0.2 |

### 6.3.2 Comparison of metagenomic datasets using CRASS algorithm

A principal concept of the CRASS algorithm is that it represents another way of reference-independent comparative analysis of metagenomic datasets. In CRASS, compared to the Compareads algorithm described above, short sequencing reads from two or more metagenomic datasets are combined into a common set of longer contigs named a cross-assembly. The CRASS algorithm determines whether every contig is shared between two or more initial samples. It is assumed that the reads that are included in individual cross-contigs are derived from the same biological source within the metagenomes being analysed. Authors of the method proposed to treat the obtained information about the shared contigs by different distance formulae in order to assess

relatedness between the initial metagenomic datasets (Dutilh et al. 2012). The formulae are based on both presence/absence (qualitative comparison) of contigs and abundance of contigs assessed by a number of reads included in the contigs (quantitative comparison).

Qualitative comparison of AP-, SH- and WGA-based datasets using the SHOT formula (Dutilh et al. 2012) is demonstrated in Figure 6.4 showing the comparison of the metagenomic datasets at both sequencing depth levels of 40,000 (Figure 6.4A) and 4,000 reads (Figure 6.4B). In both cases the results show clear clustering of AP-based datasets into two distinct clusters according to their sampling sites. WGA_A datasets also formed separate clusters in both cases, whereas SH_A, SH_B and WGA_B failed at correct site-specific clustering.

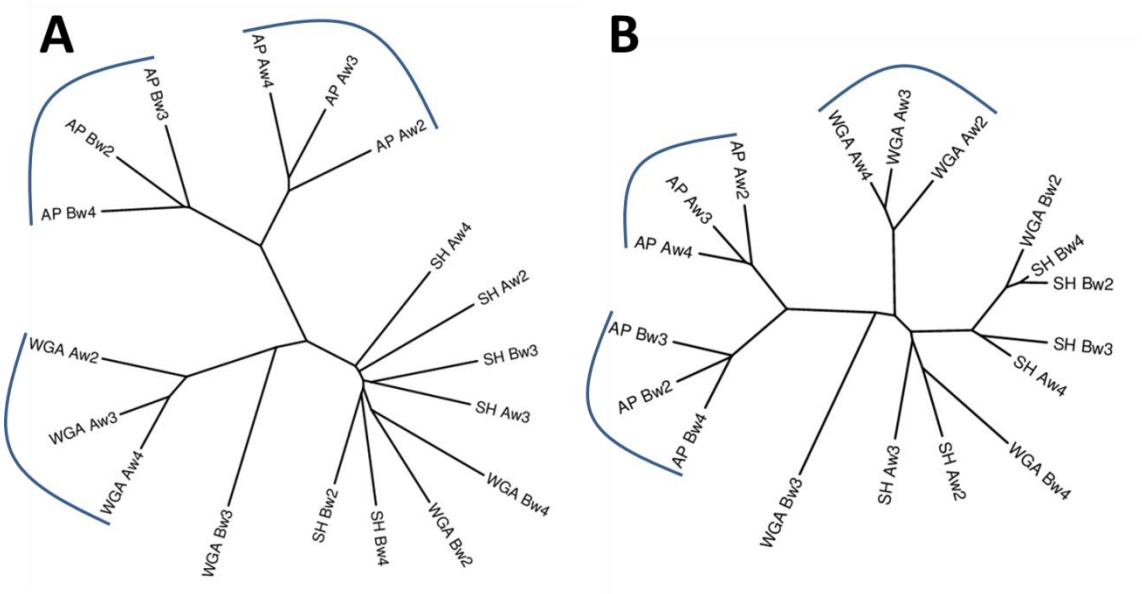



**Figure 6.4. Cladograms representing a comparison of metagenomic datasets obtained by different sequencing approaches using the CRASS algorithm and the SHOT formula for distances calculations.** (**A**) The comparison performed using 40 000 read datasets. (**B**) the comparison performed using 4000 read datasets.

Surprisingly quantitative comparison of the datasets using the READS formula revealed correct clustering not only for AP-based datasets but also for WGA-based ones however only for datasets having 40,000 reads (Figure 6.5A). At 4,000 reads (Figure 6.5B) only AP-based and WGA_A datasets were clustered correctly according to the origin of the corresponding soil samples.

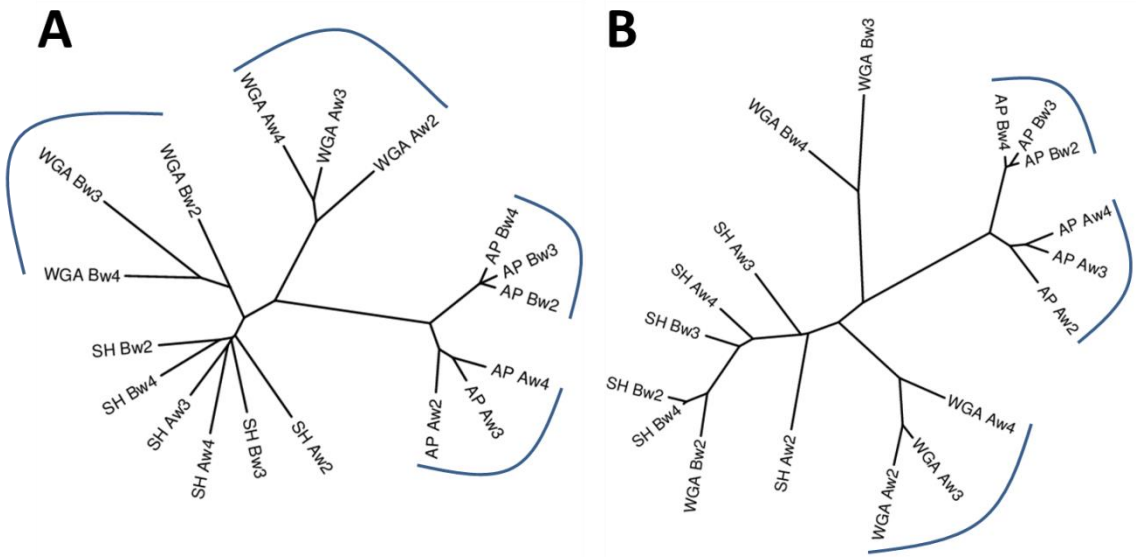



**Figure 6.5. Cladograms are representing a comparison of metagenomic datasets obtained by different sequencing approaches using CRASS algorithm and the READS formula for distances calculating.** (**A**) The comparison performed using 40,000 read datasets. (**B**) the comparison performed using 4,000 read datasets.

## 6.4 Conclusions

Analysis of the metagenomic datasets using reference-independent comparative metagenomics approaches was performed in this chapter. Results presented again confirmed that AP-based sequencing allows for the generation of site-specific DNA profiles from metagenomes of visually similar soils. Subsequent comparison of the profiles using this method led to the reliable differentiation of the soils. It is important to note that correct site-specific discrimination of soils was achieved even using AP-based datasets having a very low number of reads (4,000). This is in contrast to the SH- and WGA-based datasets that did not show site-specific soil discrimination. However during the WGA-based dataset analysis with CRASS software conditions were found that allowed for correct separation of WGA-based datasets according to their soil sample origin. This finding confirms the previously highlighted value of the unannotated fraction of metagenomic reads that might be very significant for the comparison of the metagenomic datasets. The result was obtained only for WGA-based datasets randomly subsampled down to 40,000 reads. This indicates that increasing the number of reads for WGA and potentially for SH sequencing (i.e. the use of deep sequencing) has promise in generating profiles that will allow differentiating metagenomes from similar soil types.

# Chapter 7. Final discussion and conclusion

Soil has been viewed in the forensic community as a valuable source of evidence since the beginning of the last century for linking a suspect or objects to a crime scene or victim. Due to the success of metagenomics approaches used in ecological research it has become more realistic to use the complex soil biota composition in comparative analysis for forensic purposes. A literature review on the current methods for forensic soil analysis is presented in Chapter 1.

This final chapter discusses the major findings of the thesis and reports on the results from each of the experimental chapters within the context of the objectives of the thesis.

The research presented in this thesis evaluates the capacity of modern HTS based DNA typing approaches to be used as a new tool for forensic soil comparative analysis. Well-established metagenomic sequencing approaches; 16S rRNA sequencing, shotgun sequencing and WGA sequencing as well as a new AP-PCR-based sequencing technique were investigated.

Given that 16S and AP-PCR-based sequencing involve a PCR amplification stage, preliminary experiments on evaluation of the purity of commercial DNA polymerases assessing the bacterial DNA traces retained after production of the polymerase are described in Chapter 2. Amongst eight DNA polymerases from different manufactures tested only three showed completely clear negative controls under the conditions that are normally used for the amplification of the 16S rRNA genes with universal bacterial primers.

One of the requirements for all HTS techniques, which are the basis of whole metagenomics, is the use of high quality template DNA. In Chapter 2 five different soil

DNA extraction kits were tested on three different soil types for the DNA yield, purity and background contamination, as well as for the applicability of the resultant DNA preparations for the subsequent enzymatic reactions. Two commercial kits were successful at extracting high purity DNA suitable for downstream applications. Based on these findings future work would need to evaluate reagent quality and purity at multiple and separate laboratories to ensure no biases are introduced. A way to overcome the problem using bioinformatics approaches has been recently reported (Young et al. 2014). The authors proposed to perform sequencing of extraction blank controls or PCR blank controls along with the targeted samples followed by subsequent deduction of the sequences found in negative controls from 16S rRNA targeted sequencing datasets. This approach might be an option, but at the same time it can introduce biases since the features found in the negative controls may be the same as those found in the targeted samples and their deduction can affect the final conclusion about taxonomic composition of samples of interest. This is particularly true in cases where the amount of initial DNA is low and the stochastic effect of PCR amplification is considerably high.

The last question considered in Chapter 2 was the impact of soil storage conditions on the content of the extracted DNA preparation. There were no significant differences found after storage of soils for one month at different temperatures (ambient, +4 $^{o}$C and -20 $^{o}$C). Despite the results obtained, it would be wrong to exclude the possibility that there are some soils that are affected by storage conditions.

The same soil types which were used for preliminary extraction experiments were then used for comparison of metagenomic approaches. Two of these soils were visually similar. The results of the discrimination of the soils using 16S sequencing,

which is the most widely used method for phylogenetic analysis of the microbial communities, are presented in Chapter 3. It was demonstrated that 16S rRNA sequencing allowed for reliable differentiation of the visually different soil samples however the power of discrimination of visually similar soils was low, with a high false positive error rate of 83%.

Another PCR-based method for soil metagenome evaluation was introduced in Chapter 4. The approach represents random evaluation of the full assemblage of soil DNA based on single arbitrary primed PCR (AP-PCR based sequencing). After a number of optimisation steps, the AP-PCR-based sequencing was shown to be successful at the discrimination of both visually similar and visually different soils. This was confirmed by the likelihood ratio model used in this chapter which gave a low rate of false positives and false negatives of 19% and 0% for similar soils and 3% and 11% for contrasting soils respectively. This was a very satisfying achievement for the study since the reliable discriminating of visually similar soils is of high interest and importance for forensic investigation because the majority of crimes occur in urban areas with highly similar landscape, land use and vegetation type.

Shotgun sequencing (SH), a gold standard techniques in random whole metagenomics, and WGA-based sequencing (whole genome amplification), often used where limited amount of DNA material is available, were tested for their ability to discriminate visually similar soils (Chapter 5). No discrimination of visually similar soils using SH and WGA sequencing was achieved even after application of the various bioinformatics tools including annotation of the sequencing reads with different reference databases at all available levels of taxonomic and metabolic classification, as well as *de novo* assembling of sequencing reads. It is important to note that application

of reference independent comparative metagenomics (Chapter 6) allowed for discriminating of WGA-based metagenomic datasets according to the collection sites. Comparison of AP-PCR datasets using reference independent algorithms also revealed a clear site-specific separation of the samples.

The methods presented in this proof-of-concept study show a significant step towards possible implementation of soil discrimination using metagenomics for forensic investigation and evidence generation. The results obtained in the current study clearly show that potentially there are two PCR-based metagenomic approaches that might be introduced as tools for soil discrimination in forensic practice. Each has their own advantages and limitations. Both approaches, being PCR-based, are able to utilise minute amounts of DNA material. The PCR-based nature of these approaches brings some biases such as formation of artefact chimeric sequences and the impossibility in being able to evaluate the quantitative composition of the microbial community. As such, it is not expected that these methods adequately reflect the true picture of the soil microbial community composition. This drawback, however, could be negligible for forensic soil comparison as per "… *forensic scientists would use microbial community typing to compare soil samples rather than to fully characterise their components, artefacts are acceptable so long as they occur predictably and do not impact on the ability to make accurate comparisons.*" (Coyle 2008). Nevertheless the AP-PCR-based approach performed better at discrimination of visually similar soils showing a lower rate of false-positive and false-negative results than the 16S-based sequencing. Visually contrasting soils from different locations were distinguished reliably by both methods.

An advantage in the practical application of the 16S-based sequencing approach is that the cost for the 16S-based sequencing is lower in comparison to the AP-PCR-

based sequencing; as the latter involves an additional costly procedure of library preparation for every sample to be tested. Table 7.1 summarises some of the cost and time required for both the 16S and the AP-PCR-based sequencing approaches.

**Table 7.1. Comparison of 16S and AP-PCR based metagenomic soil DNA sequencing using Ion Torrent platform (counted for 10 samples).**

| | 16S rRNA sequencing | | | AP-PCR-based sequencing | | |
|---|---|---|---|---|---|---|
| | **Reagent** | **Cost, $AUD** | **Time, h** | **Reagent** | **Cost, $AUD** | **Time, h** |
| DNA Extraction | Zymo research ZR soil DNA Kit | 61 | 0.5 | Zymo research ZR soil DNA Kit | 61 | 0.5 |
| DNA Amplification | Qiagen, HotStar Taq | 0.5 | 2 | Qiagen, HotStar Taq | 0.5 | 2 |
| PCR primers | IDT DNA | 3.6 | n/a | IDT DNA | 0.01 | n/a |
| PCR purification | Qiagen, QIAquick Gel Extraction | 48 | 1.5 | Qiagen, QIAquick PCR Purification | 40 | 0.5 |
| Concentration measurements | Life technologies, QuBit dsDNA HS assay | 13.7 | 0.25 | Life technologies, QuBit dsDNA HS assay | 13.7 | 0.25 |
| Library preparation | n/a | | | Life Technologies, IonExpress Library Prep Kit | 2000 | 1 day |
| Ion Torrent Sequensing | Life Technologies, 318 Chip | 1450 | 1 day | Life Technologies, 318 Chip | 1450 | 1 day |
| | | | | | | |
| Total | | 1576.8 | 2 days | | 3565.2 | 3 days |

Both techniques allow for the use of a barcoding procedure which in turn makes feasible sequencing of multiple samples in one run. Careful purification and measurement of the concentration of the amplification products are common requirements for both methods.

Bioinformatic analysis used in the current study for the 16S- and AP-PCR-based datasets are the most widely used pipelines in microbial ecology for these types of data. As expected the selection of reference databases for annotation may affect the resulting conclusion, since no common comprehensive and unbiased database has been

developed so far. For example annotation of 16S sequencing datasets is limited by selection of appropriated databases from Greengenes (DeSantis et al. 2006) and RDP (Ribosomal Database Project) (Cole et al. 2014) that have a high degree of overlap. Annotation of AP-PCR datasets is more flexible since it can be performed by using any database available, including both nucleotide and protein databases, producing both taxonomic and metabolic profiles of the soil samples. Variety and suitability of the reference databases for AP-PCR sequencing annotation might improve detection of different characteristic features of metagenomes that in turn could result in better discrimination of soils. Moreover, different reference-independent approaches can also be used for the comparison of AP-PCR-based datasets.

It is evident that there is a lot of unexplored potential behind both approaches. Targeted sequencing approach requires more genetic markers to be evaluated for better soil specimens discrimination and the development of multiplex amplification. For the AP-PCR method it is an examination of the mechanism of the AP-PCR amplification and the evaluation of the effect of both the primer sequence and composition on subsequent discriminating of the samples.

By increasing the number of samples analysed from each location, and also by increasing the number of distinct geographical locations, it will become possible to improve the proposed LR model, as the power of discrimination of these sequencing approaches is related to the number of samples taken. The large scale investigation of temporal microbial variation would further strengthen any tool that is developed. As the sample sizes increase, the tool will move from the model developed in this study to one that has sufficient power as a useful investigative tool and ultimately to a method that can be presented in court. For presentation in a court of law the development of a

sufficient sample size and distinct geographic profiles will need to be supported with a determination of the limitations of the method, including false positive and negative rates. This can be achieved by validation of the method being developed via blind trials, mock case work and a period of casework hardening in order to achieve the levels required for acceptance.

# Appendix A to Chapter 2

## Evaluation of soil DNA extraction, amplification and storage impacts on the soil microbial community DNA typing

Soil DNA extracts are named according to the DNA extraction kit used, sampling sites and the number of replicative extract. Thus names of DNA specimens extracted with Zymo Research ZR soil DNA extraction kit start with capital letters 'ZR', MoBio Power Soil DNA extraction kit – 'PS'; the next capital letter represents sampling location such as 'A' for location A (Flinders University), 'B' for location B (Warradale Reserve) and 'C' for location C (Brighton Esplanade); a digit (from 1 to 5) represent a number of replicate taken from the site. Thus, 'ZR_A_1' means that DNA was extracted using ZR DNA extraction kit from soils sample taken from location A and has replicate number one.

For storage experiments DNA extracts' names also contain abbreviations explaining a way the soil sample had been treated before the extraction occurred. Thus suffix 'initial' means that the DNA extraction was performed straight after soil sampling; 'IPA' shows if washing with 2-Propanol was performed; '+4', '-20' and 'RT' shows temperature of soil sample storage at $+4^{o}C$, $-20^{o}C$ and ambient temperature respectively; 'ii' or 'iv' show soil sample storage period of 2 or 4 weeks respectively. For example, 'A_IPA_RT_ii' means that the soil sample was collected from location A, washed with 2-Propanol and sored at ambient temperature for two weeks.

**Table A1. Average relative intensities of the peaks found in LH-PCR profiles obtained from ZRsoil DNA extracts.**

| PCR fragment size, bp | ZR_A_1 | ZR_A_2 | ZR_A_3 | ZR_A_4 | ZR_A_5 | ZR_B_1 | ZR_B_2 | ZR_B_3 | ZR_B_4 | ZR_B_5 | ZR_C_1 | ZR_C_2 | ZR_C_3 | ZR_C_4 | ZR_C_5 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **311** | 8.7 | 10.9 | 13.9 | 7.5 | 14.4 | 12.4 | 11.5 | 11.8 | 11.6 | 12.6 | 3.5 | 3.6 | 3.4 | 3.6 | 3.4 |
| **312** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **314** | 49.9 | 49.5 | 53.1 | 58.4 | 53.6 | 21.0 | 19.7 | 19.8 | 23.9 | 22.1 | 46.6 | 49.8 | 50.6 | 50.3 | 52.7 |
| **316** | 9.3 | 9.5 | 0.0 | 0.0 | 0.0 | 7.4 | 8.1 | 8.5 | 7.1 | 7.7 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **325** | 13.0 | 14.4 | 13.9 | 14.3 | 13.4 | 4.8 | 4.8 | 3.7 | 4.3 | 3.9 | 11.8 | 12.2 | 12.9 | 12.3 | 11.9 |
| **328** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.8 | 0.6 | 0.9 | 0.7 | 0.8 | 0.5 | 0.5 | 0.4 | 0.5 | 0.4 |
| **331** | 2.0 | 1.1 | 1.2 | 2.4 | 1.4 | 2.9 | 2.3 | 3.0 | 3.1 | 3.6 | 1.8 | 1.8 | 1.7 | 1.6 | 1.4 |
| **332** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.6 | 0.6 | 0.5 | 0.6 | 0.4 | 1.4 | 1.2 | 1.2 | 1.2 | 1.1 |
| **339** | 6.1 | 9.4 | 8.3 | 5.2 | 9.3 | 18.8 | 20.2 | 16.4 | 17.2 | 16.2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **340** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 9.6 | 8.1 | 8.4 | 9.2 | 8.7 | 11.0 | 9.4 | 9.5 | 9.1 | 9.9 |
| **341** | 3.2 | 1.0 | 0.7 | 3.0 | 2.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 11.2 | 9.5 | 9.0 | 9.7 | 9.5 |
| **342** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 2.0 | 2.0 | 3.1 | 2.5 | 1.5 | 3.3 | 3.6 | 3.1 | 3.4 | 2.7 |
| **343** | 1.7 | 1.1 | 0.9 | 1.8 | 0.7 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **344** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.7 | 0.9 | 1.1 | 0.6 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **345** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.7 | 0.8 | 0.6 | 0.8 | 0.5 |
| **346** | 4.0 | 2.4 | 1.8 | 5.3 | 3.6 | 10.6 | 12.0 | 13.3 | 10.6 | 12.4 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **347** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.3 | 1.7 | 1.6 | 1.5 | 1.2 |
| **348** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **349** | 1.4 | 0.0 | 0.7 | 1.8 | 0.5 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **350** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 4.3 | 5.1 | 5.2 | 4.5 | 6.3 | 3.4 | 3.4 | 3.8 | 3.5 | 3.3 |
| **351** | 0.6 | 0.7 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **352** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 3.6 | 2.5 | 2.3 | 2.4 | 2.0 |
| **353** | 0.3 | 0.0 | 0.0 | 0.0 | 0.0 | 4.0 | 4.1 | 4.3 | 4.1 | 3.4 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |

**Table A2. Average relative intensities of the peaks found in LH-PCR profiles obtained from PowerSoil DNA extracts.**

| PCR fragment size, bp | PS_A_1 | PS_A_2 | PS_A_3 | PS_A_4 | PS_A_5 | PS_B_1 | PS_B_2 | PS_B_3 | PS_B_4 | PS_B_5 | PS_C_1 | PS_C_2 | PS_C_3 | PS_C_4 | PS_C_5 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **311** | 7.1 | 7.8 | 7.4 | 6.2 | 7.8 | 6.5 | 6.1 | 6.9 | 5.8 | 5.7 | 3.3 | 2.8 | 3.1 | 3.1 | 2.9 |
| **312** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.1 | 1.3 | 2.0 | 1.0 | 1.2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **314** | 42.9 | 45.1 | 45.1 | 44.8 | 45.0 | 28.3 | 27.5 | 26.0 | 28.5 | 27.1 | 54.2 | 53.4 | 54.0 | 48.5 | 49.6 |
| **316** | 4.0 | 3.5 | 4.0 | 3.4 | 3.4 | 7.0 | 6.3 | 7.5 | 6.8 | 6.3 | 2.7 | 2.5 | 2.3 | 3.1 | 2.8 |
| **325** | 5.7 | 8.7 | 4.7 | 5.8 | 4.5 | 2.4 | 2.4 | 3.1 | 2.5 | 2.4 | 7.9 | 9.4 | 9.3 | 11.4 | 10.1 |
| **328** | 0.4 | 0.3 | 0.4 | 0.5 | 0.3 | 1.3 | 1.1 | 0.9 | 1.1 | 0.9 | 0.6 | 0.5 | 0.4 | 0.5 | 0.5 |
| **331** | 2.2 | 2.0 | 2.2 | 2.3 | 1.9 | 2.2 | 2.0 | 1.8 | 2.2 | 1.9 | 2.3 | 2.3 | 2.5 | 2.1 | 2.5 |
| **332** | 0.8 | 0.7 | 0.9 | 0.9 | 0.9 | 1.3 | 1.1 | 0.9 | 1.1 | 0.9 | 1.7 | 1.5 | 1.7 | 1.4 | 1.4 |
| **339** | 6.2 | 8.2 | 4.5 | 5.5 | 4.5 | 12.3 | 14.1 | 14.4 | 14.5 | 12.9 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **340** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 11.8 | 12.4 | 12.8 | 11.8 | 14.7 | 5.6 | 6.4 | 6.8 | 6.2 | 6.8 |
| **341** | 6.8 | 5.1 | 6.2 | 5.8 | 6.3 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 8.0 | 8.4 | 8.0 | 8.3 | 12.7 |
| **342** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.9 | 2.5 | 3.0 | 2.9 | 2.7 | 5.6 | 5.1 | 4.7 | 6.2 | 4.7 |
| **343** | 1.6 | 0.7 | 1.5 | 1.8 | 2.2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **344** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 5.4 | 5.0 | 4.8 | 5.2 | 5.7 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **345** | 2.8 | 4.2 | 2.5 | 2.5 | 2.5 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.6 | 0.5 | 0.7 | 0.8 | 0.0 |
| **346** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 9.1 | 9.3 | 8.0 | 8.4 | 9.3 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **347** | 10.7 | 8.0 | 11.5 | 11.6 | 12.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 3.4 | 3.1 | 2.5 | 2.9 | 1.9 |
| **348** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **349** | 8.8 | 5.7 | 9.1 | 9.1 | 8.7 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **350** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 3.3 | 3.0 | 2.6 | 2.7 | 2.7 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **351** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 2.8 | 3.0 | 2.6 | 3.5 | 2.8 |
| **352** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **353** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 6.2 | 5.9 | 5.3 | 5.6 | 5.7 | 1.2 | 1.1 | 1.3 | 2.0 | 1.3 |

**Table A3. Average relative intensities of the peaks found in LH-PCR profiles obtained from soil A stored at different conditions.**

| PCR fragment size, bp | A_initial | A_initial _ IPA | A_RT _ii | A_IPA_RT _ ii | A_+4 _ii | A_IPA_ +4 _ii | A_-20 _ii | A_IPA_ -20 _ii | A_RT _iv | A_IPA_RT _iv | A+4 _iv | A_-20 _iv | A_IPA_-20 _iv |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **311** | 14.8 | 14.9 | 14.5 | 20.2 | 17.5 | 17.2 | 16.1 | 17.0 | 18.3 | 17.0 | 23.1 | 20.9 | 21.6 |
| **312** | 0.0 | 0.0 | 4.1 | 0.0 | 8.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **314** | 39.7 | 42.7 | 43.4 | 46.7 | 46.1 | 39.4 | 41.5 | 40.1 | 42.0 | 39.2 | 18.9 | 41.1 | 46.2 |
| **316** | 1.6 | 1.2 | 0.0 | 1.4 | 0.0 | 0.9 | 0.0 | 1.2 | 0.9 | 2.8 | 3.0 | 2.0 | 2.4 |
| **325** | 10.7 | 11.4 | 13.4 | 10.4 | 12.6 | 12.7 | 12.3 | 12.2 | 11.3 | 6.9 | 14.8 | 11.3 | 10.1 |
| **328** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **331** | 1.7 | 1.7 | 1.6 | 1.2 | 1.3 | 1.8 | 1.8 | 1.5 | 1.8 | 2.1 | 2.2 | 1.4 | 1.0 |
| **332** | 0.5 | 0.5 | 0.5 | 0.3 | 0.0 | 0.4 | 0.6 | 0.5 | 0.9 | 1.3 | 1.1 | 0.0 | 0.0 |
| **339** | 9.7 | 8.3 | 5.7 | 7.1 | 5.0 | 8.1 | 6.0 | 8.3 | 5.0 | 5.5 | 7.9 | 8.7 | 7.3 |
| **340** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **341** | 4.4 | 3.6 | 3.1 | 3.7 | 1.5 | 4.1 | 4.2 | 4.0 | 3.7 | 6.5 | 5.5 | 4.4 | 3.3 |
| **342** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **343** | 1.9 | 2.0 | 0.0 | 1.3 | 0.0 | 1.9 | 1.7 | 1.9 | 2.1 | 2.6 | 2.8 | 1.5 | 1.3 |
| **344** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **345** | 0.0 | 0.0 | 9.3 | 0.0 | 5.8 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **346** | 11.6 | 10.7 | 0.0 | 5.9 | 0.0 | 10.9 | 13.7 | 10.6 | 11.5 | 12.8 | 17.4 | 6.4 | 4.9 |
| **347** | 0.0 | 0.0 | 2.7 | 0.0 | 0.9 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **348** | 0.0 | 0.0 | 0.3 | 0.0 | 0.3 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **349** | 0.5 | 0.7 | 0.0 | 0.5 | 0.0 | 0.6 | 0.7 | 0.7 | 0.8 | 1.3 | 1.1 | 0.5 | 0.5 |
| **350** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **351** | 0.0 | 0.2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.3 | 0.0 | 0.0 | 0.0 |
| **352** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **353** | 3.0 | 2.0 | 1.5 | 1.3 | 1.1 | 2.0 | 1.6 | 1.8 | 1.6 | 1.8 | 2.3 | 1.8 | 1.2 |

**Table A4. Average relative intensities of the peaks found in LH-PCR profiles obtained from soil B stored at different conditions.**

| PCR roduct size, bp | B_initial | B_initial _IPA | B_RT_i i | B_IPA_RT _ii | B_+4 _iv | B_ IPA_+4 _ii | B_-20 _ii | B_ IPA_-20 _ii | B_RT _iv | B_IPA_RT _iv | B_+4 _iv | B_IPA_+4 _iv | B_- 20_iv | B_IPA_-20 _iv |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **311** | 12.1 | 12.6 | 13.4 | 12.8 | 11.7 | 11.8 | 12.2 | 11.6 | 12.1 | 12.3 | 11.0 | 12.7 | 12.2 | 11.4 |
| **312** | 3.6 | 4.5 | 3.8 | 3.1 | 3.5 | 4.4 | 4.9 | 3.7 | 3.9 | 2.3 | 3.0 | 5.7 | 3.7 | 3.8 |
| **314** | 23.4 | 25.7 | 22.6 | 27.3 | 27.6 | 27.8 | 23.3 | 24.5 | 22.6 | 31.8 | 24.4 | 26.5 | 25.9 | 25.8 |
| **316** | 16.4 | 13.4 | 15.9 | 13.9 | 15.1 | 11.4 | 15.6 | 14.2 | 19.7 | 12.6 | 16.9 | 19.8 | 22.9 | 13.7 |
| **325** | 2.7 | 2.9 | 2.5 | 2.7 | 2.4 | 2.9 | 2.1 | 2.0 | 2.5 | 1.9 | 2.8 | 2.1 | 2.6 | 2.3 |
| **328** | 0.5 | 0.6 | 0.1 | 0.4 | 0.7 | 0.7 | 0.5 | 1.4 | 0.7 | 1.4 | 0.0 | 0.5 | 0.4 | 0.5 |
| **331** | 2.1 | 2.4 | 2.2 | 2.7 | 2.5 | 2.5 | 2.1 | 2.3 | 2.1 | 3.3 | 1.9 | 1.4 | 1.8 | 1.3 |
| **332** | 1.9 | 1.7 | 1.8 | 2.1 | 1.7 | 1.8 | 1.7 | 1.4 | 1.9 | 1.9 | 1.7 | 1.1 | 1.7 | 1.1 |
| **339** | 7.1 | 12.3 | 8.1 | 9.6 | 8.4 | 12.4 | 6.4 | 15.6 | 9.1 | 7.6 | 7.9 | 13.2 | 5.2 | 13.6 |
| **340** | 9.3 | 5.7 | 3.5 | 6.4 | 5.6 | 4.7 | 7.2 | 5.4 | 4.9 | 4.5 | 7.9 | 5.4 | 5.9 | 7.0 |
| **341** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **342** | 3.3 | 3.0 | 3.6 | 3.6 | 5.2 | 3.4 | 5.4 | 4.4 | 3.9 | 4.9 | 4.4 | 4.7 | 2.0 | 8.0 |
| **343** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **344** | 2.3 | 0.6 | 4.7 | 1.1 | 2.4 | 0.2 | 2.5 | 0.3 | 1.4 | 1.2 | 1.6 | 0.0 | 2.7 | 0.0 |
| **345** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **346** | 9.4 | 9.3 | 10.7 | 9.6 | 7.6 | 10.4 | 9.3 | 8.1 | 8.7 | 9.6 | 9.6 | 4.0 | 8.2 | 6.7 |
| **347** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **348** | 0.0 | 0.2 | 0.0 | 0.2 | 0.0 | 0.4 | 0.0 | 0.6 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **349** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **350** | 1.7 | 2.1 | 3.9 | 2.0 | 1.6 | 2.1 | 2.0 | 1.9 | 2.7 | 1.1 | 2.0 | 0.8 | 1.5 | 1.6 |
| **351** | 0.8 | 0.0 | 0.0 | 0.0 | 0.7 | 0.2 | 0.8 | 0.4 | 0.7 | 0.7 | 1.4 | 0.7 | 1.1 | 1.0 |
| **352** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **353** | 3.3 | 3.0 | 3.3 | 2.6 | 3.4 | 2.9 | 4.1 | 2.2 | 3.4 | 2.8 | 3.5 | 1.6 | 3.4 | 2.1 |

**Table A5. Average relative intensities of the peaks found in LH-PCR profiles obtained from soil C stored at different conditions.**

| PCR fragment size, bp | Sample location, storage, treatment and replica number | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | C_initial | C_initial _IPA | C_RT _ii | C_IPA_RT _ii | C_+4 _ii | C_IPA_+4 _ii | C_-20 _ii | C_IPA_-20 _ii | C_RT _iv | C_IPA_RT _iv | C_IPA_+4 _iv | C_IPA_-20 _iv |
| **311** | 3.9 | 2.7 | 4.5 | 4.7 | 4.7 | 3.5 | 4.7 | 5.2 | 4.4 | 4.9 | 4.9 | 5.9 |
| **312** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **314** | 26.9 | 23.7 | 23.3 | 27.4 | 28.5 | 25.9 | 26.2 | 27.7 | 23.9 | 25.1 | 24.7 | 25.8 |
| **316** | 11.1 | 9.1 | 6.9 | 9.8 | 10.4 | 10.1 | 9.3 | 10.7 | 9.7 | 9.0 | 8.8 | 9.8 |
| **325** | 11.3 | 6.4 | 8.5 | 8.8 | 9.0 | 8.3 | 10.7 | 9.0 | 8.7 | 8.1 | 7.2 | 7.9 |
| **328** | 1.3 | 0.8 | 0.6 | 1.3 | 0.0 | 1.7 | 1.7 | 1.9 | 1.9 | 2.1 | 1.1 | 1.5 |
| **331** | 4.6 | 8.5 | 7.2 | 5.5 | 6.3 | 4.3 | 6.3 | 4.9 | 3.9 | 4.1 | 4.4 | 4.5 |
| **332** | 1.4 | 0.3 | 2.2 | 2.0 | 2.9 | 2.0 | 3.4 | 2.1 | 0.0 | 0.0 | 0.0 | 0.0 |
| **339** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **340** | 3.3 | 2.9 | 1.2 | 5.5 | 4.4 | 4.0 | 1.5 | 3.5 | 4.2 | 3.6 | 7.0 | 5.3 |
| **341** | 11.8 | 12.7 | 12.1 | 9.8 | 8.0 | 10.4 | 12.8 | 10.3 | 10.7 | 9.2 | 11.2 | 9.4 |
| **342** | 9.5 | 24.2 | 21.1 | 15.8 | 10.6 | 9.7 | 14.3 | 14.1 | 11.3 | 10.0 | 10.1 | 9.7 |
| **343** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **344** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **345** | 9.0 | 0.0 | 0.0 | 0.0 | 9.1 | 10.5 | 0.0 | 0.0 | 11.0 | 10.4 | 10.5 | 9.8 |
| **346** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **347** | 2.0 | 3.8 | 5.7 | 3.8 | 1.4 | 2.4 | 4.5 | 2.9 | 2.5 | 3.2 | 3.6 | 3.3 |
| **348** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **349** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **350** | 6.8 | 4.6 | 5.6 | 4.9 | 5.3 | 6.7 | 4.4 | 6.1 | 6.6 | 9.4 | 5.5 | 6.2 |
| **351** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **352** | 0.0 | 0.2 | 1.0 | 0.7 | 0.0 | 0.0 | 0.0 | 0.8 | 1.2 | 0.7 | 0.7 | 0.9 |
| **353** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.7 | 0.0 | 0.1 | 0.4 | 0.0 |

# Appendix B to Chapter 4

## Arbitrary primed PCR based sequencing of soil metagenome for forensic soil discrimination

Table B1 illustrates a Bray-Curtis pair-wise similarity matrix for all possible comparisons of soil sample profiles at the species level of taxonomy. Primer Seq5-RC (P5) was used for the profile generation. The matrix is then converted into a column chart with both within a site and between sites Bray-Curtis similarities score distributions (Figure 17, black and red bars, respectively).

**Table B1. Bray-Curtis pair-wise similarity matrix of the comparison of all taxonomic profiles (species level) generated with primer P5 by AP-PCR based sequencing at AGRF.**

| | AP_Aw1 | AP_Aw2 | AP_Aw3 | AP_Aw4 | AP_Asp | AP_As | AP_Aa | AP_Bw1 | AP_Bw2 | AP_Bw3 | AP_Bw4 | AP_Bsp | AP_Bs | AP_Ba | AP_Cw1 | AP_Cw2 | AP_Cw3 | AP_Cw4 | AP_Csp | AP_Cs | AP_Ca |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **AP_Aw1** | | | | | | | | | | | | | | | | | | | | | |
| **AP_Aw2** | 70 | | | | | | | | | | | | | | | | | | | | |
| **AP_Aw3** | 68 | 71 | | | | | | | | | | | | | | | | | | | |
| **AP_Aw4** | 68 | 71 | 72 | | | | | | | | | | | | | | | | | | |
| **AP_Asp** | 65 | 67 | 70 | 75 | | | | | | | | | | | | | | | | | |
| **AP_As** | 61 | 64 | 68 | 69 | 68 | | | | | | | | | | | | | | | | |
| **AP_Aa** | 66 | 67 | 70 | 74 | 72 | 70 | | | | | | | | | | | | | | | |
| **AP_Bw1** | 51 | 56 | 50 | 53 | 50 | 50 | 51 | | | | | | | | | | | | | | |
| **AP_Bw2** | 58 | 63 | 56 | 57 | 55 | 55 | 55 | 66 | | | | | | | | | | | | | |
| **AP_Bw3** | 56 | 61 | 55 | 56 | 53 | 53 | 53 | 67 | 71 | | | | | | | | | | | | |
| **AP_Bw4** | 56 | 61 | 53 | 54 | 51 | 51 | 53 | 71 | 68 | 68 | | | | | | | | | | | |
| **AP_Bsp** | 54 | 59 | 53 | 55 | 54 | 53 | 54 | 69 | 67 | 65 | 73 | | | | | | | | | | |
| **AP_Bs** | 47 | 50 | 45 | 47 | 46 | 47 | 46 | 61 | 61 | 61 | 61 | 62 | | | | | | | | | |
| **AP_Ba** | 52 | 56 | 50 | 52 | 51 | 52 | 51 | 66 | 66 | 63 | 68 | 68 | 66 | | | | | | | | |
| **AP_Cw1** | 53 | 55 | 53 | 54 | 52 | 53 | 55 | 45 | 49 | 47 | 46 | 49 | 42 | 45 | | | | | | | |
| **AP_Cw2** | 57 | 58 | 59 | 58 | 58 | 54 | 58 | 46 | 50 | 48 | 47 | 49 | 41 | 45 | 60 | | | | | | |
| **AP_Cw3** | 52 | 54 | 54 | 55 | 54 | 50 | 54 | 43 | 47 | 46 | 45 | 47 | 40 | 43 | 55 | 58 | | | | | |
| **AP_Cw4** | 57 | 57 | 57 | 58 | 58 | 53 | 58 | 48 | 51 | 49 | 48 | 50 | 42 | 46 | 60 | 70 | 60 | | | | |
| **AP_Csp** | 51 | 52 | 53 | 57 | 58 | 49 | 56 | 43 | 46 | 44 | 45 | 47 | 41 | 44 | 49 | 53 | 53 | 58 | | | |
| **AP_Cs** | 55 | 55 | 56 | 58 | 58 | 53 | 59 | 46 | 51 | 49 | 48 | 50 | 44 | 47 | 52 | 59 | 56 | 62 | 64 | | |
| **AP_Ca** | 56 | 55 | 56 | 57 | 58 | 52 | 58 | 47 | 50 | 48 | 49 | 50 | 43 | 47 | 52 | 61 | 54 | 64 | 63 | 69 | |

The average similarity scores for each group were then transformed to a model of the normal distribution of Bray-Curtis similarity scores, shown on Figure B1.

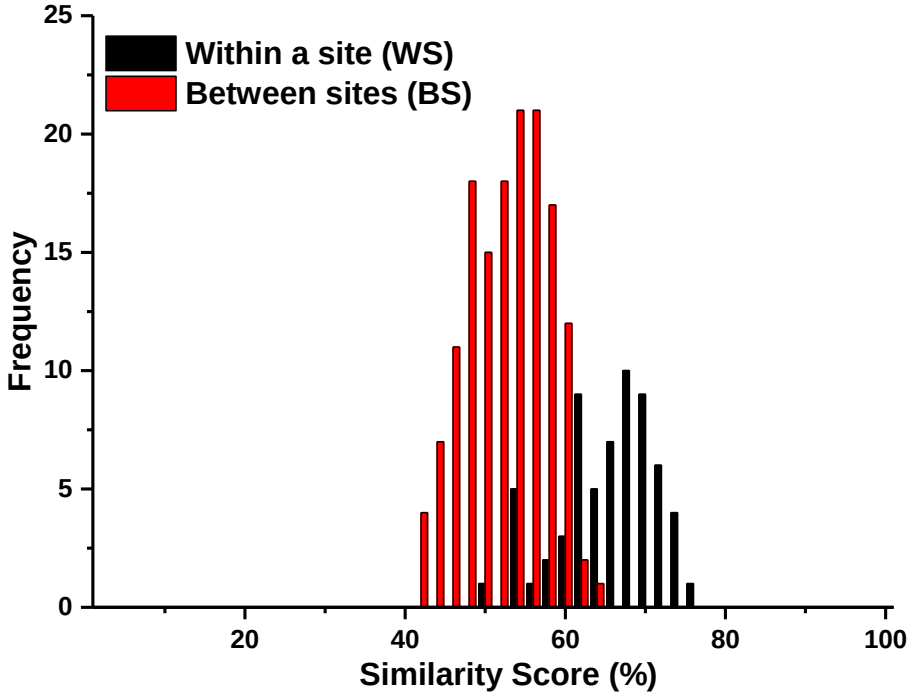



**Figure B1. Bray-Curtis similarity scores distribution for within site group (black bars) and between sites group (red bars).**

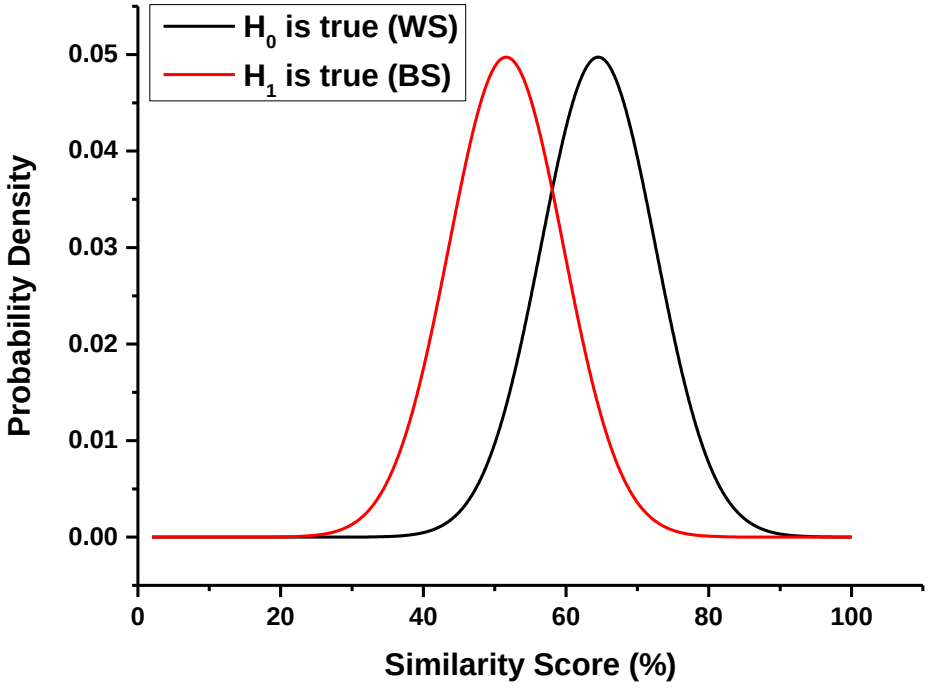



**Figure B2. Gaussian distribution of Bray-Curtis similarity scores for the within site group (black lines) and between sites group (red lines).** H0 is the hypothesis that the two samples have the same origin, while the H1 is the hypothesis that they have a different origin

**Table B2. $Log_{10}$(LR) values derived from Bray-Curtis similarity scores.**

| **Comparison of visually similar soils A & B** | | | | **Comparison of different soils A & B & C** | | | |
|---|---|---|---|---|---|---|---|
| Within a site A-A & B-B | | Between sites, A-B | | Within a site A-A & B-B & C-C | | Between sites A-C & B-C | |
| Similarity Score, % | $Log_{10}$(LR) | Similarity Score, % | $Log_{10}$(LR) | Similarity Score, % | $Log_{10}$(LR) | Similarity Score, % | $Log_{10}$(LR) |
| 70 | 1.06 | 51 | -0.63 | 70 | 1.06 | 53 | -0.45 |
| 68 | 0.88 | 58 | -0.01 | 68 | 0.88 | 57 | -0.09 |
| 68 | 0.88 | 56 | -0.18 | 68 | 0.88 | 52 | -0.54 |
| 71 | 1.15 | 56 | -0.18 | 71 | 1.15 | 57 | -0.09 |
| 71 | 1.15 | 56 | -0.18 | 71 | 1.15 | 55 | -0.27 |
| 72 | 1.24 | 63 | 0.44 | 72 | 1.24 | 58 | -0.01 |
| 66 | 0.71 | 61 | 0.26 | 66 | 0.71 | 54 | -0.36 |
| 67 | 0.80 | 61 | 0.26 | 67 | 0.80 | 57 | -0.09 |
| 71 | 1.15 | 50 | -0.72 | 71 | 1.15 | 53 | -0.45 |
| 71 | 1.15 | 56 | -0.18 | 71 | 1.15 | 59 | 0.08 |
| 68 | 0.88 | 55 | -0.27 | 68 | 0.88 | 54 | -0.36 |
| 68 | 0.88 | 53 | -0.45 | 68 | 0.88 | 57 | -0.09 |
| | | 53 | -0.45 | 60 | 0.17 | 54 | -0.36 |
| | | 57 | -0.09 | 55 | -0.27 | 58 | -0.01 |
| | | 56 | -0.18 | 60 | 0.17 | 55 | -0.27 |
| | | 54 | -0.36 | 58 | -0.01 | 58 | -0.01 |
| | | | | 70 | 1.06 | 45 | -1.16 |
| | | | | 60 | 0.17 | 46 | -1.07 |
| | | | | | | 43 | -1.34 |
| | | | | | | 48 | -0.89 |
| | | | | | | 49 | -0.81 |
| | | | | | | 50 | -0.72 |
| | | | | | | 47 | -0.98 |
| | | | | | | 51 | -0.63 |
| | | | | | | 47 | -0.98 |
| | | | | | | 48 | -0.89 |
| | | | | | | 46 | -1.07 |
| | | | | | | 49 | -0.81 |
| | | | | | | 46 | -1.07 |
| | | | | | | 47 | -0.98 |
| | | | | | | 45 | -1.16 |
| | | | | | | 48 | -0.89 |

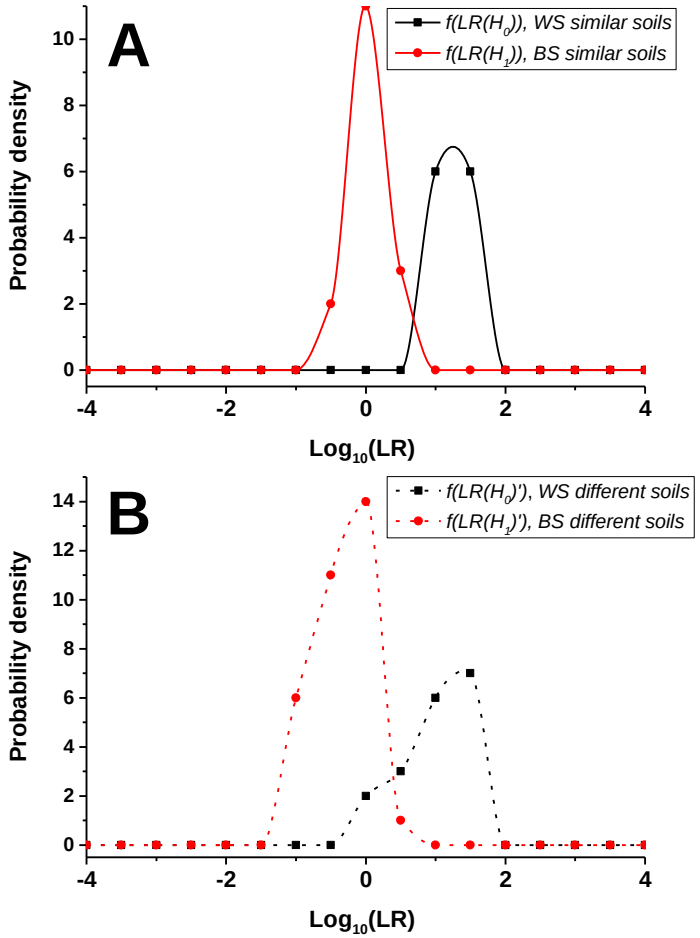



**Figure B3. Estimated probability density functions (PDFs) of $Log_{10}$(LR) values for (A) visually similar soils, (B) different soils.** Functions *f(LR(H0))* and *f(LR(H1))* represent the WS probability density function and BS probability density function for similar looking soils samples. Functions *f(LR(H0'))* and *f(LR(H1)')* represent the WS probability density function and BS probability density function for different looking soils samples.

The following publication (Khodakova, 2013) describes the application of AP-PCR based sequencing of soils A, B and C using SEED database for profile annotation.

Forensic Science International: Genetics Supplement Series 4 (2013) e39–e40

Contents lists available at ScienceDirect

Forensic Science International: Genetics Supplement Series

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

# Forensic analysis of soils using single arbitrarily primed amplification and high throughput sequencing

Anastasia S. Khodakova [a,*], Leigh Burgoyne [a], Damien Abarno [b], Adrian Linacre [a]

[a] School of Biological Sciences, Flinders University, Adelaide, South Australia, Australia
[b] Forensic Science South Australia, Adelaide, South Australia, Australia

ARTICLE INFO

*Article history:*
Received 31 August 2013
Accepted 2 October 2013

*Keywords:*
Forensic soil analysis
Soil metagenomics
High throughput sequencing
Arbitrarily primed PCR

ABSTRACT

Soil is a remarkably complex, diverse, ubiquitous, and easily transferred material which can reveal highly useful information to assist forensic investigations. In spite of its potential usefulness, the use of genetic soil analysis appears to be currently underestimated in forensic practice. Herein we report on the use of single arbitrarily primed amplification followed by high throughput sequencing of DNA fragments for the comparison of soil samples. The composition and functional attributes of soil microbial communities from three different locations were compared and shown to be different based on the metagenomic sequencing data obtained.



## 1. Introduction

Soil is the one of the largest reservoirs of genetic diversity. Discrimination of soils by traditional DNA fingerprinting techniques based on variation in fragment length has been found unreliable due to a lack of reproducibility. The recent advances in high throughput sequencing (HTS), including sequence specific, shotgun and whole genome sequencing, has advanced metagenomic research and soil microbiome analysis. HTS generates large amounts of genetic data which creates challenges in data storage, quality control and analysis [1]. Appropriately generated and analyzed metagenomic data can be extremely valuable for life science research as well as in assisting forensic investigations. Present research describes an application of arbitrarily primed polymerase chain reaction as a selective pre-enrichment soil DNA amplification stage for subsequent HTS. Our data illustrates the capability to identify dominant features in soil community structure that are well-known and ubiquitous in soils [2] and allowed for the successful discrimination between three different soils sampled at different locations.

## 2. Materials and methods

The three soil sites, located 3–4 km apart in the Adelaide, South Australia, (A: S35.029006 E138.571508, B: S35.016136 E138.536675, C: S35.021317 E138.515922) were sampled. Total genomic DNA was isolated from 0.25 g of each soil sample using the ZR Soil Microbe DNA MiniPrep kit (Zymo Research). PCR amplification was performed using the following reaction mixture (25 μL): 0.4 μM of the single arbitrary primer with sequence 5′-GGAGGTGGTGTTCGAGGG-3′, 2.5 mM $Mg^{2+}$, 0.2 mM of each dNTPS, 0.5 U HotstarTaq DNA polymerase (Qiagen), 1× HotstarTaq buffer (Qiagen) and 1–5 ng of the extracted soil DNA as a template. PCR amplification regime of 95 °C for 15 min, 42 cycles of 94 °C for 30 s, 55 °C for 30 s, 72 °C for 60 s and a final extension of 72 °C for 7 min was used. Sequencing was performed by the ACRF Cancer Genomics Facility on an ion PGM sequencer (Life Technologies) on an Ion 314 chip using barcoded adapters. Sequences were then annotated on the MG-RAST online software [3]. Similarity search between the obtained reads and the SEED database [4] was processed with a minimum alignment length of 15 bases and an *E*-value cut-off of $10^{-5}$. All compared distributions were normalized as a function of the number of annotated sequences for each metagenome. Data of functional and taxonomic distributions were then statistically analyzed using the STAMP software [5]. Fisher's exact test was performed and taxa with $p$-values $< 0.05$ (labelled with an asterisk on the plots) were considered to be significantly different between the different metagenomes.

## 3. Results and discussion

Metagenomic DNA from three different locations was amplified and analyzed by means of HTS. We utilized an arbitrarily primed PCR as a method for sequence independent amplification, selection and pre-enrichment of the metagenomic DNA. A total of 449,262

* Corresponding author. Tel.: +61 8 8201 5003; fax: +61 8 8201 2905.
*E-mail addresses:* anastasia.khodakova@flinders.edu.au, askhodakova@gmail.com (A.S. Khodakova).

1875-1768/$ – see front matter © 2013 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.fsigss.2013.10.019

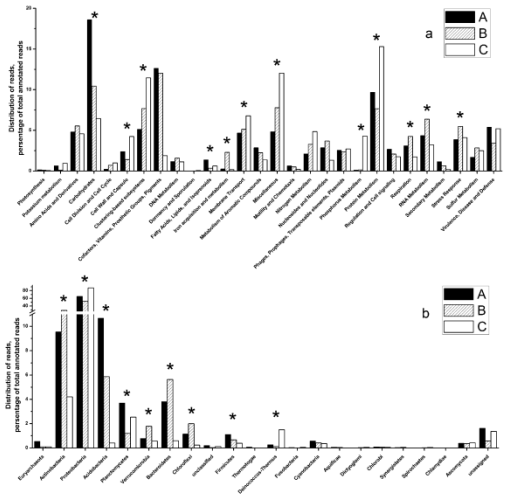



**Fig. 1.** The relative distributions of major metabolic classes (a) and taxonomic groups (Phylum) (b) in the three soil metagenomes. Asterisks indicate those categories with significantly different abundance in soils ($p < 0.05$).

reads were generated with an average length of $189 \pm 49$ bp. Approximately $30 \pm 2\%$ of the reads were annotated with the SEED protein database using MG-RAST (MG-RAST ID #: 4518019.3 (sample A), 4518020.3 (sample B), 4518019.3 (sample C)). The differences between the soils were evident from the comparison of the relative abundances of functional genes classified at the lowest level of resolution (Fig. 1a). Thus metabolic subsystems: carbohydrates, fatty acids, lipids and isoprenoids were prevalent in sample A; iron acquisition and metabolism, RNA metabolism, respiration and stress response were prevalent in sample B; and clustering-based subsystems, membrane transport, miscellaneous, cell wall and capsule, protein metabolism, and phosphorous metabolism were prevalent in sample C. Comparison of taxonomic profiles also revealed specific features that strongly differentiated the soil metagenomes. The different taxonomic patterns (Fig. 1b) were due to difference in the abundance of major taxonomic groups. The Acidobacteria, Planctomicetes and Firmicutes were more abundant in sample A, Actinobacteria, Bacteroidetes, Chloroflexi and Verrucomicrobia were more abundant in sample B and Proteobacteria and Deinococcus-Thermus were more abundant in sample C. Taxonomic and functional profiles of the obtained metagenomic data were dominated by the similar features that are known to be abundant and ubiquitous in soils, as per 'gold-standard' methods such as 16 S rRNA and shotgun sequencing [6]. The proposed approach has demonstrated potential for site-specific soil discrimination between different locations, highlighting the potential of metagenomic profiling to be used in forensic comparison of soils. Further research is required to gain a better understanding of variation across different spatial scales.

## 4. Conclusion

The relative distributions of major metabolic classes (a) and taxonomic groups (Phylum) (b) in the three soil metagenomes. Asterisks indicate those categories with significantly different abundance in soils ($p < 0.05$).

Site-specific soil profiles generated by the proposed amplification technique and analyzed by HTS showed an applicability of arbitrarily primed amplification for generation of specific DNA profiles of metagenomic DNA samples. The sequence independent and multiple targeted mechanism of the arbitrarily primed amplification permitted analysis of metagenomic DNA samples by both taxonomic annotation and major metabolic classes identification. Both annotation methods resulted in successful discrimination of soil samples taken from three different locations. Further research is needed with a larger number of metagenomic samples across different habitats and soils in order to evaluate the reproducibility and performance of the approach in comparison with other high throughput sequencing technologies.

## Role of funding

Funding was provided by the Department of Justice South Australia.

## Conflict of interest

The authors declare no competing interest.

## References

[1] M.B. Scholz, C.-C. Lo, P.S.G. Chain, Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis, Current Opinion in Biotechnology 23 (2012) 9–15.
[2] P. Janssen, Identifying the dominant soil bacterial taxa in libraries of 16S rRNA and 16S rRNA genes, Applied and Environmental Microbiology 72 (2006).
[3] F. Meyer, D. Paarmann, M. D'Souza, R. Olson, E.M. Glass, M. Kubal, et al., The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes, BMC Bioinformatics 9 (2008) 386.
[4] R. Overbeek, T. Begley, R.M. Butler, J.V. Choudhuri, H.-Y. Chuang, M. Cohoon, et al., The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes, Nucleic Acids Research 33 (2005) 5691–5702.
[5] D.H. Parks, R.G. Beiko, Identifying biologically relevant differences between metagenomic communities, Bioinformatics (Oxford, England) 26 (2010) 715–721.
[6] N. Fierer, J.W. Leff, B.J. Adams, U.N. Nielsen, S.T. Bates, C.L. Lauber, et al., Cross-biome metagenomic analyses of soil microbial communities and their functional attributes, Proceedings of the National Academy of Sciences of the United States of America 109 (2012) 21390–21395.

# Appendix C to Chapter 5

## Random whole metagenomics as a tool for forensic soil discrimination

OPEN ACCESS Freely available online

PLOS ONE

# Random Whole Metagenomic Sequencing for Forensic Discrimination of Soils

**Anastasia S. Khodakova[1]*, Renee J. Smith[1], Leigh Burgoyne[1], Damien Abarno[1,2], Adrian Linacre[1]**

**1** School of Biological Sciences, Flinders University, Adelaide, Australia, **2** Forensic Science South Australia, Adelaide, Australia

## Abstract

Here we assess the ability of random whole metagenomic sequencing approaches to discriminate between similar soils from two geographically distinct urban sites for application in forensic science. Repeat samples from two parklands in residential areas separated by approximately 3 km were collected and the DNA was extracted. Shotgun, whole genome amplification (WGA) and single arbitrarily primed DNA amplification (AP-PCR) based sequencing techniques were then used to generate soil metagenomic profiles. Full and subsampled metagenomic datasets were then annotated against M5NR/M5RNA (taxonomic classification) and SEED Subsystems (metabolic classification) databases. Further comparative analyses were performed using a number of statistical tools including: hierarchical agglomerative clustering (CLUSTER); similarity profile analysis (SIMPROF); non-metric multidimensional scaling (NMDS); and canonical analysis of principal coordinates (CAP) at all major levels of taxonomic and metabolic classification. Our data showed that shotgun and WGA-based approaches generated highly similar metagenomic profiles for the soil samples such that the soil samples could not be distinguished accurately. An AP-PCR based approach was shown to be successful at obtaining reproducible site-specific metagenomic DNA profiles, which in turn were employed for successful discrimination of visually similar soil samples collected from two different locations.

**Citation:** Khodakova AS, Smith RJ, Burgoyne L, Abarno D, Linacre A (2014) Random Whole Metagenomic Sequencing for Forensic Discrimination of Soils. PLoS ONE 9(8): e104996. doi:10.1371/journal.pone.0104996

**Editor:** Carles Lalueza-Fox, Institut de Biologia Evolutiva - Universitat Pompeu Fabra, Spain

**Received** May 16, 2014; **Accepted** July 15, 2014; **Published** August 11, 2014



**Data Availability:** The authors confirm that all data underlying the findings are fully available without restriction. All metagenomic data are available from the MG-RAST database (http://metagenomics.anl.gov/).

**Funding:** Funding for this research was provided by the Attorney General's Office of South Australia. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

* Email: anastasia.khodakova@flinders.edu.au

## Introduction

Soil can be found on items submitted for forensic analysis, however there is currently no reliable method to compare the DNA content of soils for forensic purposes. Soil, owing to its inherent features, adheres under fingernails, to cars, tools, weapons or items of clothing and can transfer during the commission of a criminal act [1]. Soil can also be useful associative evidence in the investigation of wildlife crimes, such as poaching. The presence of soil is often recorded during the forensic examination of exhibits. Due to the lack of a validated analytical method, or set of techniques for meaningful comparison of soil samples, this evidential type provides only limited value in investigations. There is therefore a need to develop such comparative methodologies.

Traditionally forensic analysis of soils involves comparison of its chemical-physical and biological properties [2]. Over the past decades many studies have been undertaken utilizing the chemical profiles of soil using a wide variety of novel sophisticated and rapid analytical methods such as, FTIR [3], X-ray [4] and elemental analysis [5,6]. These methods are mainly mineralogical techniques and define geological characteristics of soil, which differ across a regional scale. Therefore these techniques may be unable to discriminate soils within a small locality [7]. The potential for discriminating soils at a local scale exists with methods of soil microbial community analysis that have been applied for forensic purposes [8,9]. Previous attempts at DNA based analysis of soils used DNA fingerprinting techniques which evaluate fragment length variation such as terminal restriction fragment length polymorphism (TRFLP) [7,10], denaturing gradient gel electrophoresis (DGGE) [11], amplified ribosomal DNA restriction analysis (ARDRA) [12] and length heterogeneity-polymerase chain reaction (LH-PCR) [5]. Many fragments in the resultant DNA fingerprint appear identical in length but differ in sequence leading to erroneous conclusions of similarity that would be avoided if the DNA sequences of the fragments were known. These methods are rapid and permit high throughput analysis but have insufficient resolution to discriminate complex soil mixtures [13]. All these methods have potential for use in forensic comparisons, however a lack of reproducibility and the potential for false inclusions has restricted their implementation in a forensic setting.

Development of new platforms for high-throughput DNA sequencing (HTS) has made it more affordable and led to the significant growth of HTS-based studies [14 16]. The application of HTS to soil science has allowed for new insight on the diversity of soil microbial communities inhabiting various biomes [17 19].

Gene-targeted, or locus-specific, sequencing which typically targets the 16S rRNA gene is used for characterization of the taxonomic composition and diversity of microbial communities [20,21]. Shotgun sequencing is primarily a method for studying

the functional structure of the communities which aims to examine the entire genetic assemblage and, being amplification-independent, relies on variation and commonality of the collective genomes found in a given environmental sample [22,23]. Shotgun typing allows for a more comprehensive perspective on the whole microbial community but is limited by its propensity to favour identification of the most dominant members over rarer organisms [24]. In order to access the rare species found in such a complex matrix as soil, ultra-deep DNA sequencing is required [25].

Soil samples obtained during forensic investigations, by their nature, put specific requirements on any metagenomic approach. The samples are often small and sufficient amount of the sample should remain after analysis for independent re-testing if required. Soil DNA extraction procedure, as an initial step of metagenomic analysis, should provide high quality DNA with a good yield. Commercially available soil DNA extraction kits is an preferable option offering forensic investigators a means for standardizing soil DNA extraction [26 28]. Gene-targeted sequencing based on PCR amplification technique is able to analyse the minute amounts of template DNA recovered in forensic samples. The need for a relatively large amount of initial DNA template for shotgun sequencing makes this approach less suitable for forensic oriented metagenomic analysis but whole-genome amplification (WGA), using Phi 29 DNA polymerase, represents an effective way of enabling whole-genome shotgun sequencing from small quantities of DNA [29].

The ability to identify DNA from the entire genetic composition of a complex soil mixture is desirable for forensic investigation as the DNA from a wide range of organisms may be present: these include the DNA from bacteria, fungi, nematodes, mammals, plant material, and from insect remains. These can be used to generate a rich DNA profile for comparison and meaningful discrimination between samples. Targeted metagenomics technique that are limited to one particular locus, such as the 16S (small subunit (SSU) of rRNA in prokaryotes [30]), ITS (internal transcribed spacers widely used for fungi [31]), or 18S (nuclear SSU rRNA a widely used phylogenetic marker in eukaryotes [32]), do not detect the variability of the entire soil biota thus providing less information for comparison and differentiation of soil samples.

Metagenomic sequencing approaches have been reported that can reliably differentiate soil microbial communities from different soil types and different land use [17,33]. However from a forensic point of view, discrimination of visually similar soil samples taken from geographically different urban areas (community parks with similar plant cover, residential suburbs, soils of similar land management) is of greater importance [7,34].

We report on the assessment of the ability of random whole metagenomic sequencing approaches to produce reproducible site-specific DNA profiles that can be employed for comparative analysis and discrimination between soils of different locality for application in forensic science. We show an assessment of shotgun and WGA-based sequencing techniques as well as the use of a single arbitrarily primed DNA amplification (AP-PCR) for metagenomic soil DNA analysis [35]. The use of AP-PCR was first reported in the 1990s [36,37] and has been applied to genotyping [38,39] and the study of microbial communities [40].

## Materials and Methods

### Soil sampling

Soil samples were collected from two different sites in Adelaide in July 2013; Location A (S35 01 43.42 E138 34 16.26) and Location B (S35 00 58.09 E138 32 12.03). These locations are separated by approximately 3 km. For each site, triplicate samples

**Table 1.** General characteristics of full sequencing data.

| Sequencing approach | Average number of reads (range) | Number of Mbp | Average read length, bp ± SD | Failed QC(%) | Number of reads with predicted protein coding regions (range) | Number of reads with predicted rRNA genes (range) | Number of assigned features to M5NR database (%) | Number of assigned features to SEED Subsystems (%) | Number of assigned features to M5RNA database (%) |
|---|---|---|---|---|---|---|---|---|---|
| **Shotgun** | 672 542 (531 108–806 483) | 133.6 | 197±73 | 20 | 464 929 (325 410–582 708) | 82 151 (62 899–96 886) | 35 | 43 | 1.3 |
| **APPCR** | 468 187 (74370–1 074 266) | 70.7 | 142±69 | 24 | 287 840 (49 902–617 609) | 44 896 (5 868–104 247) | 26 | 35 | 0.0 |
| **WGA** | 911 553 (506 028–2 012 359) | 178.5 | 198±75 | 20 | 549 355 (354 930–1 032 625) | 96 117 (61 694–187 539) | 26 | 30 | 0.8 |

Statistical data represented as mean ± Standard Deviation (SD). Percentage of sequences matching to the M5NR, M5RNA and SEED Subsystems databases was determined with an E-value cut-off of $E<1\times10^{-5}$. QC = quality control.

doi:10.1371/journal.pone.0104996.t001

were taken 1 m apart from the upper 1 cm of the soil layer. The samples collected from Location A and Location B represented a dark loam rich in organic matter and were visually very similar. No specific permits were required for these locations and activities. The field studies did not involve endangered or protected species. The soil samples were placed in individual sterile plastic tubes and stored at −20°C until analysis. DNA extraction was performed within 24 hours of soil sampling.

## DNA extraction, amplification and sequencing

Metagenomic DNA was isolated from 50 mg of each soil sample using the ZR Soil Microbe DNA Kit (Zymo Research, USA) following the manufacturer's recommendations. The quality of the DNA extracts was verified by gel electrophoresis in a 1% agarose gel stained with ethidium bromide. DNA concentrations were determined using a Qubit dsDNA HS Assay Kit (Invitrogen, USA) on a Qubit fluorometer (Life technologies, USA).

For each of the six samples, WGA was conducted with 20 ng DNA using Phi29 DNA polymerase (REPLI-g, Qiagen, Germany). The quality of amplification products was determined by 1% agarose gel electrophoresis and by quantification on a Qubit fluorometer (Life technologies, USA) after purification with a QIAquick PCR Kit (Qiagen, Germany).

Amplification of extracted soil DNA was performed with an arbitrarily chosen oligonucleotide primer (sequence 5′-GGAGGTGGTGTTCGAGGG-3′), previously reported for generating soil DNA fingerprints [41]. As a template, 4 ng of metagenomic DNA was used. The 25 µL final reaction volume contained 1×Hotstar Taq buffer (Qiagen, Germany), 2.5 mM $Mg^{2+}$, 0.2 mM of each dNTPs, 0.4 µM of the arbitrary chosen primer, and 0.5 U HotstarTaq DNA polymerase (Qiagen, Germany). An initial 15 min denaturation step at 95 °C was followed by 42 cycles of 30 s at 94°C, 30 s at 55°C and 1 min at 72°C. A final extension step of 7 min at 72°C was used to complete the reaction. The quality and concentration of purified PCR products (QIAquick PCR Kit, Qiagen, Germany) were determined as described for WGA procedure.

All the manipulations were performed in dedicated DNA extraction and PCR-mixing hoods using sterile DNA/RNA free water (Ambion, USA) and DNA/RNA free plasticware (Eppendorf, Germany). All the procedures of the extraction and amplification were conducted with the necessary no-template controls, including extraction blank controls.

Library preparation from 100 ng of the corresponding DNA specimen for all three methods under evaluation followed by sequencing was performed at the Australian Genome Research Facility (AGRF, http://www.agrf.org.au/, Adelaide, SA, Australia) using Ion Torrent technology (Ion Torrent PGM Sequencer; Life Technologies, USA) on a separate Ion 318 chip for each of the sequencing approaches.

## Processing of sequencing data

Raw sequence datasets were uploaded to the Metagenome Rapid Annotation using Subsystem Technology (MG-RAST) server (http://metagenomics.nmpdr.org/) (Meyer et al., 2008) and filtered from low-quality reads prior to annotation. Metagenomic datasets were annotated to protein genes against the M5NR database and SEED Subsystems database resulting in protein-derived taxonomic and metabolic profiles, respectively. In addition taxonomic profiles were generated by comparison of the metagenomic datasets with the M5RNA ribosomal database also available in MG-RAST. The MG-RAST default annotation parameters such as maximum E-value $<1\times10^{-5}$, minimum length of alignment of 15 bp, and minimum sequence identity of 60%, were used to identify the best database matches. Metagenomic profiles were generated at all available MG-RAST taxonomic (phylum to species) and metabolic (level 1 to functions) levels of hierarchy. To adjust the differences in sequencing effort across samples, two common procedure of standardization were taken:

1. In the first approach metagenomic profiles were generated using full datasets of the high-quality reads obtained for each sample. For the metagenomic profiles comparison the relative abundance scores for each taxon and metabolic feature were determined by the percentages of respective reads over the total assigned reads. In the text the relative abundance scores found both for the taxonomic and metabolic features are represented as an average ± SD (standard deviation) across all datasets (if not mentioned otherwise).
2. A second approach was based on comparison of metagenomic profiles generated from randomly subsampled datasets of 49 000 annotated reads per sample.

Metagenomic datasets are freely available on the MG-RAST web-server (http://metagenomics.anl.gov/). The MG-RAST sample IDs are listed in the Table S1.

## Statistical analysis of data

The species richness was estimated by rarefaction analysis preformed in MG-RAST. The analysis was performed for total taxa identified with the M5NR protein database in randomly subsampled metagenomic datasets (including Bacteria, Archaea, Eukaryota, Viruses, unclassified and other sequences).

Statistical comparison of metagenomes was conducted on square root transformed data using the statistical package Primer v.6 for Windows (Version 6.1.13, PRIMER-E, Plymouth) [43]. To assess the similarity of the taxonomic and metabolic compositions between soil samples, the Bray-Curtis pair-wise similarity measure was employed. The resulting Bray-Curtis similarity matrices were then used for hierarchical agglomerative clustering (CLUSTER) with the results displayed as group average dendrograms. Similarity profile analysis (SIMPROF) was used to test for multivariate structure in the clusters formed. Non-metric multi-dimensional scaling (NMDS) of Bray-Curtis similarities was performed as an unconstrained ordination method to graphically visualise inter-sample relationships. The program RELATE in the PRIMER package was used to calculate the Spearman rank correlation between Bray-Curtis similarity matrices generated from differently standardised datasets at the same level of taxonomic or metabolic resolution [44].

Metagenome profiles were further analysed using canonical analysis of principal coordinates (CAP) using the PERMANOVA+ version 1.0.3 3 add-on to PRIMER [45] as a constrained ordination method to test for significant differences among the *a priori* groups in multivariate space. All metagenomic profiles were divided into 6 groups according to the sequencing approach applied and origin of the samples. The *a priori* hypothesis of 'no difference' within groups was tested at both taxonomic and metabolic levels using CAP analysis by evaluation of a *P*-value obtained after 9999 permutations. The strength of the association between multivariate data and the hypothesis of group differences was indicated by the value of the squared canonical correlation ($\delta_1^2$). An appropriate number of principal coordinates axes ($m$) used for the CAP analysis were chosen automatically by the CAP routine to minimize errors of a misclassification. In order to validate the ability of the CAP model to classify correctly the

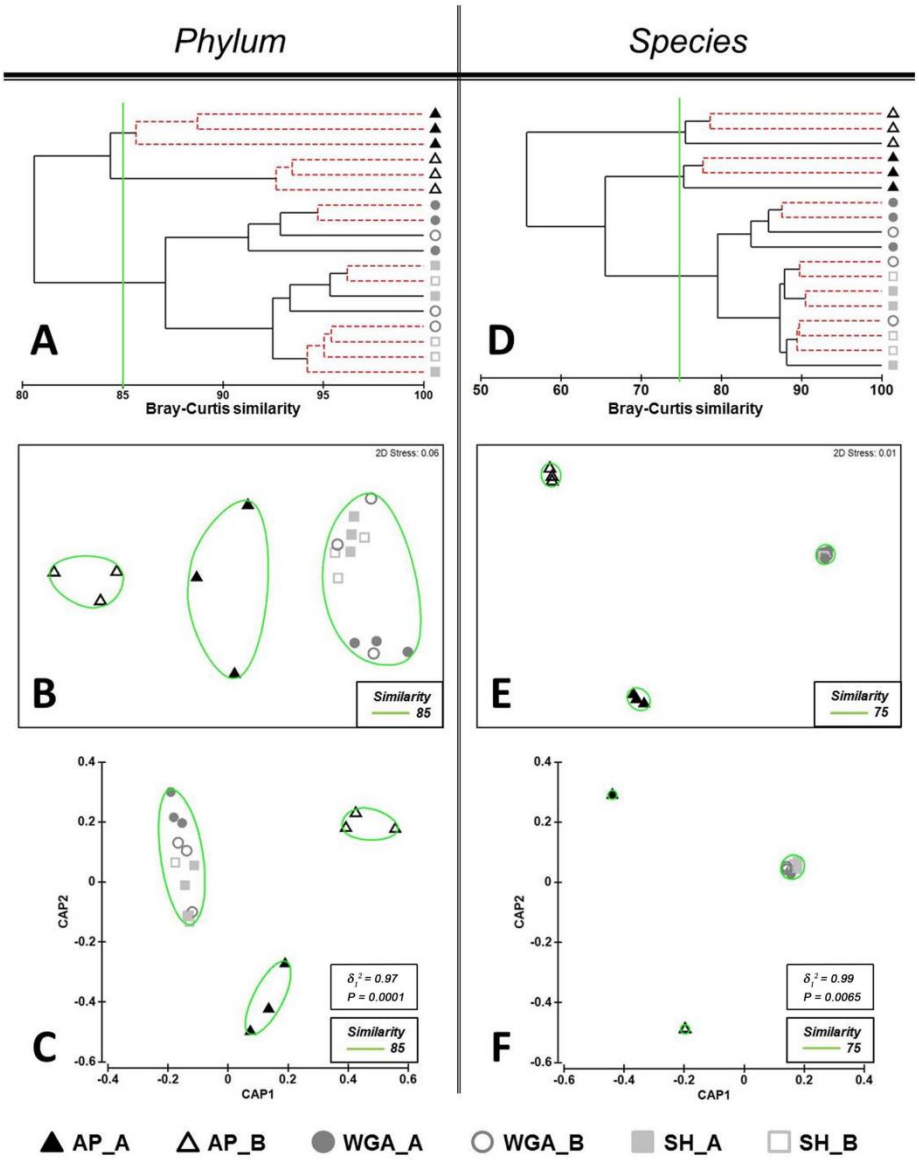



**Figure 1. Comparison of the taxonomic soil profiles generated on full datasets at the phylum (A, B, C) and species (D, E, F) resolution levels.** Bray-Curtis distance similarity matrix was calculated from the square-root transformed abundance of DNA fragments matching taxa in the M5NR database (E-value $<1\times10^{-5}$). The Bray-Curtis matrix was used for generating CLUSTER dendrogram, NMDS and CAP ordination

plots. **CLUSTER analysis (A and D).** Red dotted branches on the CLUSTER dendrogram indicate no significant difference between metagenomic profiles (supported by the SIMPROF analysis, $p<0.05$). **NMDS unconstrained ordination (B and E).** The NMDS plot displays distances between samples. Data points that are closer to each other represent samples with highly similar metagenomic profiles. **CAP constrained ordination (C and F).** CAP analysis tests for differences among the groups in multivariate space. The significance of group separation along the canonical axis is indicated by the value of the squared canonical correlation ($\delta_1^2$) and P-value. A contour line on the NMDS and CAP ordinations drawn round each of the cluster defines the superimposition of clusters from CLUSTER dendrogram at the selected level of similarity.
doi:10.1371/journal.pone.0104996.g001

samples according to their appropriate groups a cross-validation procedure was performed for the chosen value of $m$ [46].

## Results

### Notation and general characteristics of sequencing datasets

Obtained datasets were grouped and named according to their sequencing approach and soil sampling sites. Thus samples processed by the AP-PCR approach have a common prefix "AP", shotgun sequenced samples "SH", and WGA assisted sequencing "WGA". Each dataset designation identifies the location from where the sample was collected: "_A" samples collected from location A; and "_B" samples collected from location B. For example the abbreviation AP_A indicates a sample collected at location A and sequenced by the AP-PCR-based method.

For each soil DNA sample three datasets were generated from the same DNA template using three sequencing approaches. Shotgun metagenome sequencing resulted in an average of 672 542 (531 108 806 843) sequence reads with an average sequence length of 198±73 bases for a total of >133 Mbp of sequence. Sequencing datasets after WGA consisted of an average of 911 554 (506 028 2 012 359) sequences with an average of 198±75 bases in length for a total of >178 Mbp. The AP-based approach gave an average of 468 187 (74 370 1 047 266) reads with an average 143±69 bases in length for a total of >70.7 Mbp (Table 1). Datasets were annotated using the online MG-RAST server [42]. Approximately 20% of low quality reads were eliminated from each dataset at the filtering step. Only 25 35% of the reads which contained predicted protein coding regions (49 902 689 805 reads per sample), were taxonomically assigned using M5NR protein database. While 30 40% of reads assigned to the SEED Subsystems database were used for generation of metabolic profiles (Table 1). Each of the metagenomic datasets according to the MG-RAST statistics contained approximately 10% of reads with predicted rRNA gene fragments. The subsequent annotation revealed no reads from the AP-based dataset and only 1% of the reads from the SH- and WGA-based datasets matched the M5RNA database.

### Taxonomic profiling of metagenomes

The analysis of metagenomic data within MG-RAST occurs both for protein coding genes and ribosomal (rRNA) genes. And therefore analysis of taxonomy can be performed in two ways.

Taxonomic classification of protein gene fragments showed that 85 (±4)% of the annotated reads were assigned to Bacteria, with 4.5 (±2.7)% of reads also matched to Eukaryota and 0.6 (±0.4)% to Archaea. The remaining 10 (±1)% of reads were not assigned. Bacterial taxa *Proteobacteria, Actinobacteria* and *Bacteroidetes* dominated in all metagenomic datasets representing close to 70% of protein annotated reads. Additional phyla including *Chloroflexi, Planctomycetes, Acidobacteria Firmicutes, Cyanobacteria, Verrucomicrobia* represented less than 5% of reads. Among the eukaryotic taxa, *Ascomicota* was found to be the dominant microorganism 3.0 (±2.6)%. Other eukaryotic taxa such as *Streptophyta, Chordata, Basidiomycota* and *Arthropoda* collectively contributed to the remaining 1% of the annotated reads (Table S2).

Taxonomic classification of the rRNA gene fragments identified only in SH- and WGA-based datasets showed that 78 (±8)% of reads were assigned to bacterial taxa and 14.5 (±6.5)% to eukaryotic taxa (data represented as an average relative abundance of taxa between the samples of SH- and WGA-based datasets). The most abundant bacterial and eukaryotic phyla found were the same as per protein-derived taxonomic classification (described above) namely: *Actinobacteria, Proteobacteria, Bacteroidetes, Ascomycota* and *Streptophyta*. The remaining 7 (±4) % of reads were not assigned (Table S3).

Rarefaction analysis was performed on randomly subsampled metagenomic datasets (49 000 reads per sample) annotated against the M5NR non-redundant protein database. The analysis showed the differences in biodiversity (highest level of taxonomic resolution) of the datasets generated by the three metagenome sequencing approaches (Fig. S1). The SH- and WGA-based datasets demonstrated a similar numbers of identified species from location A and B. A two fold lower number of species were identified in the AP-based dataset.

### Metabolic profiling of metagenomes

Metabolic profiles for all datasets were created by matching to the SEED Subsystems database. The most abundant metabolic features found in all datasets, accounting for almost 60% of assigned reads were: clustering-based subsystems; carbohydrates; amino acids and derivatives; protein metabolism; miscellaneous; cofactors; vitamins; prosthetic groups; pigments and DNA metabolism. The relative abundance each of the remaining metabolic features represented less than 5% of reads (Table S4).

### Comparison of soil metagenome profiles based on full sequence datasets

**Comparison of protein-derived taxonomic profiles.** An initial comparison of the taxonomic structures of the metagenomes using lowest (coarsest) resolution profiles derived at the phylum level of taxonomy was performed. CLUSTER analysis with group-average linking based on Bray-Curtis similarity matrices delineated two distinct clusters with similarity of 85% formed by samples from AP-based dataset grouped according to the sites from where the samples were taken (Fig. 1A). These clusters were supported by the SIMPROF analysis that showed statistically significant ($p< 0.05$) evidence of genuine clustering, as indicated by red dotted branches on the dendrogram (Fig. 1A). Two samples from WGA_A group having 94% profiles similarity also formed such a cluster. Other samples form SH- and WGA-based datasets formed mixed clusters. For example, a sample from the WGA_B group formed a united cluster with a sample from the SH_A group and two samples from the SH_B group (similarity 94%), thus indicating that the samples from two different locations were grouped together incorrectly. One more cluster consisted of two samples from SH_A and SH_B groups with 96% of similarity.

Bray-Curtis distances between metagenomic profiles were then displayed on an NMDS plot (Fig. 1B). NMDS analysis did not reveal a clear visual separation of data. Points denoting samples

**Table 2.** Results of CAP model cross-validation of soil taxonomic profiles discrimination generated from full sequencing datasets.

| Original Group | AP_A | AP_B | WGA_A | WGA_B | SH_A | SH_B |
|---|---|---|---|---|---|---|
| Taxonomy level | ***phylum*** ($m = 6$, $\delta_1^2 = 0.97$, $P = 0.0001$) | | | | | |
| % correct | 67 | 100 | 100 | 0 | 67 | 67 |
| correct/total | 2/3 | 3/3 | 3/3 | 0/3 | 2/3 | 2/3 |
| Misclassified to group | AP_B | n/a | n/a | SH_A | SH_B | SH_A |
| | | | | SH_B | | |
| | | | | WGA_A | | |
| Taxonomy level | ***class*** ($m = 5$, $\delta_1^2 = 0.98$, $P = 0.0001$) | | | | | |
| % correct | 100 | 100 | 100 | 0 | 67 | 33 |
| correct/total | 3/3 | 3/3 | 3/3 | 0/3 | 2/3 | 1/3 |
| Misclassified to group | | | | SH_B | | SH_A |
| | n/a | n/a | n/a | WGA_A | WGA_B | WGA_B |
| | | | | SH_A | | |
| Taxonomy level | ***order*** ($m = 3$, $\delta_1^2 = 0.97$, $P = 0.0002$) | | | | | |
| % correct | 100 | 100 | 100 | 0 | 67 | 33 |
| correct/total | 3/3 | 3/3 | 3/3 | 0/3 | 2/3 | 1/3 |
| Misclassified to group | | | | SH_B | | SH_A |
| | n/a | n/a | n/a | WGA_A | SH_B | WGA_B |
| | | | | SH_A | | |
| Taxonomy level | ***family*** ($m = 10$, $\delta_1^2 = 0.99$, $P = 0.0034$) | | | | | |
| % correct | 100 | 100 | 100 | 0 | 67 | 67 |
| correct/total | 3/3 | 3/3 | 3/3 | 0/3 | 2/3 | 2/3 |
| Misclassified to group | | | | SH_B | | |
| | n/a | n/a | n/a | WGA_A | SH_B | WGA_B |
| | | | | SH_A | | |
| Taxonomy level | ***genus*** ($m = 11$, $\delta_1^2 = 0.99$, $P = 0.01$) | | | | | |
| % correct | 100 | 100 | 100 | 0 | 100 | 67 |
| correct/total | 3/3 | 3/3 | 3/3 | 0/3 | 3/3 | 2/3 |
| Misclassified to group | | | | SH_B | | |
| | n/a | n/a | n/a | WGA_A | n/a | WGA_B |
| | | | | SH_A | | |
| Taxonomy level | ***species*** ($m = 10$, $\delta_1^2 = 0.99$, $P = 0.0065$) | | | | | |
| % correct | 100 | 100 | 100 | 0 | 67 | 67 |

**Table 2.** Cont.

| Original Group | AP_A | AP_B | WGA_A | WGA_B | SH_A | SH_B |
|---|---|---|---|---|---|---|
| correct/total | 3/3 | 3/3 | 3/3 | 0/3 | 2/3 | 2/3 |
| Misclassified to group | | | | SH_B | | |
| | n/a | n/a | n/a | WGA_A | SH_B | WGA_B |
| | | | | SH_A | | |

doi:10.1371/journal.pone.0104996.t002

from WGA- and SH-based datasets were located much closer together showing a higher similarity of the profiles than points representing AP-based dataset (Fig. 1B). Overlaying clusters on the NMDS plot made visual discrimination of the patterns formed by AP-based dataset easier (Fig. 1B).

It has been noted that the distinct patterns of multi-dimensional datasets could be hidden in the low-dimensional space of NMDS ordination [46]. Consequently for the comparison of our metagenomics datasets, CAP analysis as a constrained ordination method was also performed. CAP analysis tests the hypothesis of whether there is a difference between pre-defined groups. In our research all datasets were divided into 6 groups in accordance with combined factors, including the sequencing approach applied and the origin of soil samples. The results of the CAP ordination at the phylum level demonstrated that the first squared canonical correlation was very large ($\delta_1^2 = 0.97$), indicating the significance of the CAP model. The first canonical axis showed clear separation of the samples within AP-based dataset according to the soil sampling sites. At the same time a close overlapping of the samples from the SH- and WGA-based datasets was observed (Fig. 1C). However, the cross-validation results of the CAP model for the chosen value of $m = 6$ did not confirm the above defined separation of the metagenomic datasets (Table 2). Thus, the most distinct groups, which had a 100% success under cross-validation, were AP_B and WGA_A. One sample from the AP_A group was misclassified to the AP_B group. One sample from each of the SH_A and SH_B groups were misclassified to the SH_B and SH_A groups, respectively. All the samples from the WGA_B group were misclassified to another three different groups (SH_A, SH_B and WGA_A).

It is of note that apart from AP_A and WGA_A groups at the class level of taxonomic resolution the cross-validation of the CAP model showed a 100% correct classification of the samples from AP_B group (Table 2). Additionally one sample from the SH_A group was misclassified to the WGA_B group, whereas two samples from the SH_B group were misclassified to the SH_A and WGA_B groups.

Further CLUSTER analysis, NMDS and CAP ordinations of the metagenomic samples at higher levels of taxonomy demonstrated similar patterns of differentiation as observed at the phylum and class levels (Figs.1, S2,S3). Thus, at the order, family, genus and species levels of resolution two samples from the WGA_A group and two samples from the SH_A group formed separate genuine clusters on the CLUSTER dendrograms (Fig. 1D, S2D, S3A, S3D). Two more genuine mixed clusters were observed consisting of the samples from the SH_B and the WGA_B groups. NMDS and CAP ordinations at all levels of resolution clearly displayed three distinct clusters; two clusters consisting of the samples from the AP_A and the AP_B groups and one mixed cluster of samples from all the other groups (Fig. 1, S2, S3). Cross-validation results of the CAP models at all levels of resolution, starting from the class level, showed an accurate 100% correct classification of samples from the AP-based dataset (Table 2). Despite the visual overlapping of the SH- and WGA-based data points shown on the ordination plots (Fig. 1, S2, S3), the samples from WGA_A group were classified 100% correctly across all levels of taxonomic resolution (Table 2). Of note was that, at the genus level, all samples from the SH_A group were also successfully allocated.

**Comparison of taxonomic profiles based on rRNA gene fragment classification.** Taxonomic profiles were generated only for the SH- and the WGA-based datasets where the rRNA gene fragments matched to the M5RNA database. The AP-based dataset was excluded from the consecutive comparative analysis

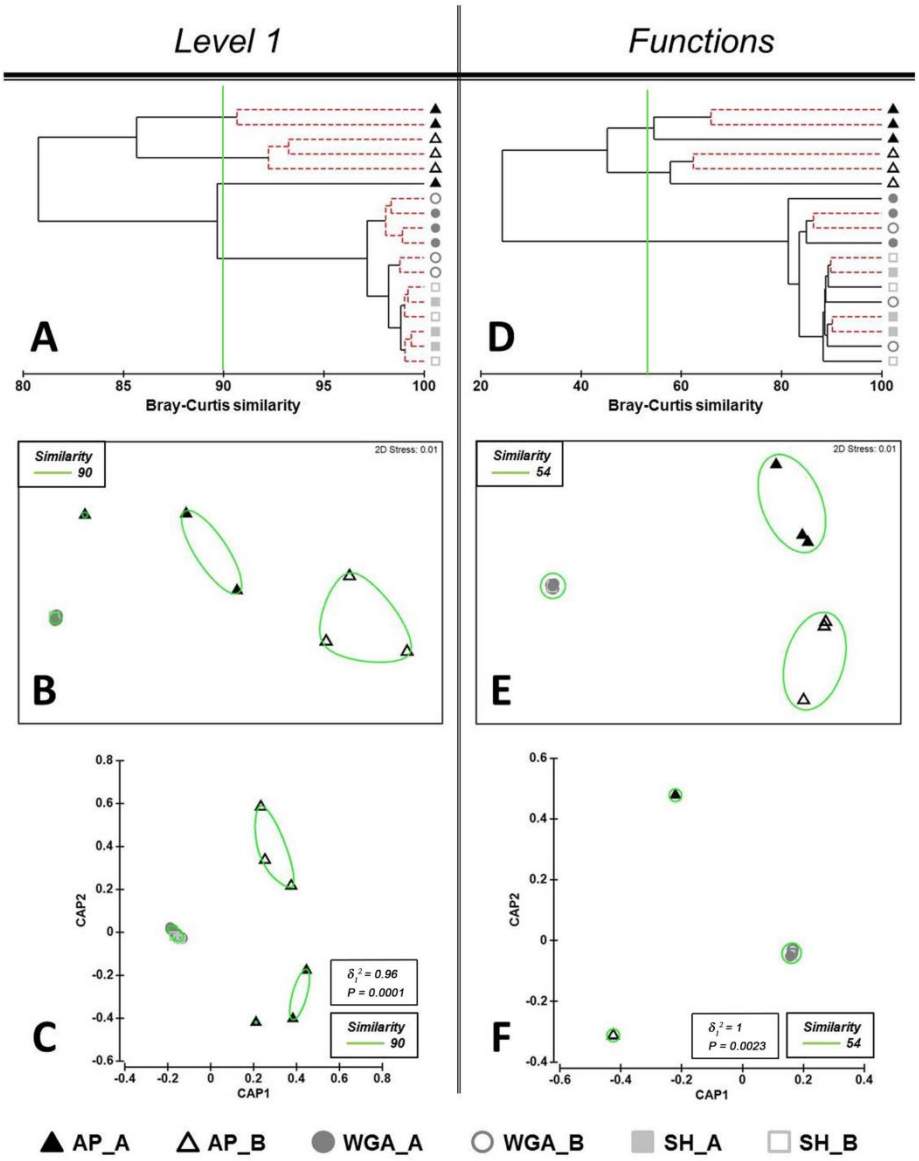



**Figure 2. Comparison of the metabolic soil profiles generated on full datasets at the subsystems level 1 (A, B, C) and subsystems function (D, E, F) resolution levels.** Bray-Curtis distance similarity matrix was calculated from the square-root transformed abundance of DNA fragments matching taxa in the SEED database (E-value $<1\times10^{-5}$). The Bray-Curtis matrix was used for generating CLUSTER dendrogram, NMDS and

CAP ordination plots. **CLUSTER analysis (A and D).** Red dotted branches on the CLUSTER dendrogram indicate no significant difference between metagenomic profiles (supported by the SIMPROF analysis, $p<0.05$). **NMDS unconstrained ordination (B and E).** The NMDS plot displays distances between samples. Data points that are closer to each other represent samples with highly similar metagenomic profiles. **CAP constrained ordination (C and F).** CAP analysis tests for differences among the groups in multivariate space. The significance of group separation along the canonical axis is indicated by the value of the squared canonical correlation ($\delta_1^2$) and P-value. A contour line on the NMDS and CAP ordinations drawn round each of the cluster defines the superimposition of clusters from CLUSTER dendrogram at the selected level of similarity.
doi:10.1371/journal.pone.0104996.g002

since no sequence matches to the ribosomal database were found. CLUSTER analysis of rRNA-based taxonomic profiles at the phylum level of resolution demonstrated the formation of four genuine clusters confirmed by SIMPROF analysis ($p<0.05$) (Fig. S4A). One cluster included three samples from the WGA_A group and one sample from the WGA_B group with similarity of 77%. A second cluster consisted of two samples from the SH_A group and one sample from the SH_B group with similarity of 85%. Two other mixed clusters were formed by the samples from different groups. The pattern formed by the samples from the WGA_A group was also seen on the NMDS and CAP plots with a 100% correct allocation which was confirmed by the results of cross-validation of the CAP model (Fig. S4B, S4C; Table S7). Two separate clusters formed by the samples from the WGA_A and the SH_A groups were observed at the higher levels of taxonomic resolution (genus and species) (Figs. S6). Observed groupings had a 100% correct allocation under cross-validation of the CAP model only at the genus level of classification (Table S7). The latter findings were in full accordance with the allocation of WGA_A and SH_A groups performed using protein-derived taxonomy (Table 2).

**Metabolic profiles comparison.** CLUSTER analysis of metabolic profiles generated by different sequencing approaches at the lowest level of resolution (level 1) showed that all three samples from the AP_B group formed a separate cluster with a similarity of 92% (Fig. 2A). Two samples from the AP_A group had a similarity of 90%. The third AP_A sample was bundled with the samples from SH- and WGA- based datasets forming a new mixed cluster. Importantly the SH- and WGA-based datasets consisting of 12 metagenomic samples formed one united mixed cluster with a similarity of 97% (Fig. 2A). NMDS and CAP ordinations also showed that all the points associated with the samples from SH- and WGA-based datasets produced a very compact cluster (Fig. 2B and Fig. 2C). However, according to a cross-validation procedure the most distinct groups with 100% allocation success were the AP-based groups and the WGA_A group, whereas misclassification errors were shown for the WGA_B, SH_A and SH_B groups (Table 3). Statistical comparisons of the metabolic profiles at higher resolution levels (level 2, level 3 and function) resulted in similar discriminating success (Fig. 2, S7). CLUSTER analysis showed correct site-specific grouping of the samples from AP-based dataset (Fig. 2D, S7A, S7D). All the profiles produced by SH- and WGA-based methods again formed a single unresolved cluster. NMDS and CAP ordinations demonstrated clear separation of three clusters (Fig. 2, S7), which was also the case for the metagenomic profiles comparison based on protein-derived taxonomy (Fig. 1, S2, S3). In both cases cross-validation results of the CAP model gave 100% correct classification of the samples from the AP_A, AP_B and WGA_A groups and misclassification errors for samples from the SH_A, SH_B and WGA_B groups (Table 2, Table 3).

### Comparison of metagenomic profiles based on randomly sub-sampled datasets

**Comparison of taxonomic profiles based on rRNA gene fragment classification.** CLUSTER analysis and NMDS ordination of rRNA-based taxonomic profiles at the phylum level of taxonomy demonstrated a heterogeneous mixed cluster of the samples from the SH- and WGA-based datasets with an average similarity of 70% (data not shown). Cap analysis showed 100% correct classification of samples from the WGA_A group and misclassification errors for samples from other groups. At the species level of resolution CLUSTER analysis also revealed a single heterogeneous mixed cluster with thea taxonomic profile similarity of approximately 25%. CAP analysis indicated a high degree of misclassification errors.

**Comparison of protein-derived taxonomic and metabolic profiles.** It has been proposed that in order to enable the comparison of metagenomes based on equal sequencing efforts, the datasets should be randomly sub-sampled to the size of the smallest sample [17,47]. The metagenomic datasets generated by shotgun, WGA-based and AP-PCR-based approaches were re-analysed by MG-RAST at an equivalent sequencing depth of 49 000 annotated reads per sample. Comparison of taxonomic and metabolic profiles generated from sub-sampled datasets at all levels of classification available within MG-RAST was performed by CLUSTER analysis, NMDS and CAP ordination.

Statistical analysis of the sub-sampled metagenomic datasets generated by three metagenome sequencing approaches yielded nearly identical estimates of the overall differences between soil microbial communities from locations A and B as those obtained using full sequence datasets (Figs. S8 S12, Tables S5 S6). This similarity was also confirmed using the RELATE programme which revealed a strong correlation between Bray-Curtis distance matrices (Spearman rank coefficient $r>0.9$, $p<0.0001$) generated on both full, and sub-sampled, datasets at all levels of taxonomic and metabolic resolution (Table 4).

## Discussion

Numerous ecological studies show that soil microbial communities differ between land uses and vegetation types [17,18,33,48,49]. The discrimination of geographically distinct urban soils with similar land management type and similar plant cover is of great forensic relevance [7,34]. If two soil samples appear very different visually then a simple exclusion can be made but more typically soils appear visually similar and currently no further action is taken. The vast majority of samples submitted for forensic investigation come from urban areas; here we include gardens, parkland and open spaces as well as built-up areas. Thus we focused our study on assessing the ability of three random whole metagenomic sequencing approaches to describe and differentiate the composition of soil microbial communities from two random parklands in 3 km apart within Adelaide residential areas. The vegetation categories of these locations appeared to be very similar, with widespread grass and trees species.

Along with standard metagenomic approaches such as shotgun and WGA, which are widely accepted as the most comprehensive sources of data for studying complex microbial communities, we evaluated AP-PCR as a method for generation of random metagenomic DNA profiles of soils that were then analysed by high throughput DNA sequencing. In this technique an arbitrary chosen oligonucleotide is used as a single primer that targets

**Table 3.** Results of CAP model cross-validation of soil metabolic profiles discrimination generated from full sequencing datasets.

| Original Group | AP_A | AP_B | WGA_A | WGA_B | SH_A | SH_B |
|---|---|---|---|---|---|---|
| Metabolic level | ***level 1*** ($m=2$, $\delta_1^2=0.96$, $P=0.0001$) | | | | | |
| % correct | 100 | 100 | 100 | 33 | 67 | 33 |
| correct/total | 3/3 | 3/3 | 3/3 | 1/3 | 2/3 | 1/3 |
| Misclassified to group | n/a | n/a | n/a | SH_A | SH_B | SH_A |
| | | | | SH_B | | WGA_B |
| Metabolic level | ***level 2*** ($m=11$, $\delta_1^2=1$, $P=0.0002$) | | | | | |
| % correct | 100 | 100 | 100 | 33 | 67 | 100 |
| correct/total | 3/3 | 3/3 | 3/3 | 1/3 | 2/3 | 3/3 |
| Misclassified to group | | | | SH_B | | |
| | n/a | n/a | n/a | WGA_A | SH_B | n/a |
| Metabolic level | ***level 3*** ($m=12$, $\delta_1^2=1$, $P=0.0009$) | | | | | |
| % correct | 100 | 100 | 100 | 33 | 67 | 67 |
| correct/total | 3/3 | 3/3 | 3/3 | 1/3 | 2/3 | 2/3 |
| Misclassified to group | | | | SH_B | | |
| | n/a | n/a | n/a | WGA_A | SH_B | SH_A |
| Metabolic level | ***functions*** ($m=10$, $\delta_1^2=1$, $P=0.0023$) | | | | | |
| % correct | 100 | 100 | 67 | 0 | 67 | 67 |
| correct/total | 3/3 | 3/3 | 2/3 | 0/3 | 2/3 | 2/3 |
| Misclassified to group | | | | SH_B | | |
| | n/a | n/a | WGA_B | WGA_A | SH_B | SH_A |
| | | | | SH_B | | |

doi:10.1371/journal.pone.0104996.t003

**Table 4.** RELATE comparison of Bray-Curtis similarity matrices.

| Taxonomic level | Spearman rank coefficient | Metabolic level | Spearman rank coefficient |
|---|---|---|---|
| phylum | 0.887 | level 1 | 0.652 |
| class | 0.944 | level 2 | 0.958 |
| order | 0.959 | level 3 | 0.967 |
| family | 0.940 | functions | 0.969 |
| genus | 0.965 | | |
| species | 0.966 | | |

The Bray-Curtis similarity matrices calculated from square root transformed abundance of DNA fragments generated based on full datasets and sub-sampled datasets.
doi:10.1371/journal.pone.0104996.t004

multiple genomic sequences producing a highly primer-sequence-specific profiles. Depending both on the primer chosen and the annealing temperature used there is sequence specific selection of complementary sequences to the primer to be DNA amplified. Based on previous studies the random amplification of polymorphic DNA, normally performed at low stringency conditions (low annealing temperature), becomes more reproducible at high stringency amplification conditions [50]. Amplification with a single arbitrary primer yields an arbitrary product pattern which might possess PCR products from both abundant species and those that are rare, again depending on the affinity of the primer.

The composition of the soil microbial communities was determined from both taxonomic classification of rRNA fragments and the taxonomic assignment of functional gene fragments. Similar taxonomic distribution of dominant microbial phyla was observed across all metagenomic datasets using these two different annotation pipelines. Reads with functional gene fragments were also used for the comparison of metagenomic datasets based on metabolic profiles.

Previous reports have indicated that comparison of metagenomes at low levels of resolution, i.e. analysis based on more broadly defined categories, results in a more conservative estimate of the distances between metagenomic profiles [17]. Low levels of taxonomic or functional classification show less overlap between samples and are therefore also used frequently for metagenomic profile discrimination [51,52]. The results of the metagenomic dataset comparison in the current study are presented at all MG-RAST taxonomic (phylum to species) and metabolic (level 1 to functions) levels of hierarchy. The comparison of metagenomes was performed with a number of unconstrained statistical tools including CLUSTER and NMDS analyses as well as constrained CAP analysis testing a predefined hypothesis that was previously shown to be successful for soil microbial communities discrimination [46,53].

SH- and WGA-based metagemonic sequencing approaches showed incorrect and inconsistent discrimination of soil samples according to sampling sites using both taxonomic (protein and ribosomal) and metabolic classifications. Comparison of the SH- and WGA-based profiles revealed not only misclassification of the samples between the locations but often between repeat analysis of each sequencing approach, with the exception of the WGA_A samples which had a 100% allocation success. The high similarity of the data generated by these methods appears to be driven by the highly similar, or even identical, dominant microorganisms found in the soil samples collected from two distinct sites of similar urban type. This supports the theory that the data generated by shotgun sequencing are commonly shifted towards describing the most abundant taxa leaving the contribution of rare microorganisms undervalued for comparative analysis [54].

A rarefaction analysis was performed to determine microbial species richness of metagenomic datasets produced by three random whole metagenome sequencing approaches for the soil samples from location A and B. The rarefaction curves computed for metagenomic datasets did not reach the plateau phase suggesting that more sequencing effort would be required to achieve species saturation. At the same time the analysis showed that the SH- and WGA-based approaches provided a higher number of species from the same number of sequence reads than the AP-based approach. The AP-PCR utilises primer dependant sequence specific selection of gene fragments and therefore unlikely to amplify all the DNA fragments present in samples. Nevertheless, despite the lower species coverage of soil metagenomes provided by the AP-based approach it allowed for a 100% correct discrimination between soils samples from different locations. This may be as a result of the pre-enrichment mechanisms of AP-PCR that are based on the primer sequence targeting both dominant and rare microorganisms equally. An AP-PCR-based strategy for whole metagenomic profile generation may be compromised by artefacts, including chimeric sequences caused by PCR amplification, which have been reported for gene-targeted (e.g. 16S) sequencing approaches [19]. It is likely that the AP-PCR based approach does not reflect the true picture of the soil microbial community composition. However, we found consistent evidence that an AP-PCR-based whole metagenome sequencing approach was able to discriminate similar soil samples based on differences in both taxonomic and metabolic compositions.

## Conclusion

In the research presented here we investigated the ability of whole metagenome analysis techniques to discriminate soil samples of similar land use and vegetation type but collected from different geographical locations. There is currently no agreed evaluation approach leading to an accurate picture of the soil metagenome structure as the true soil microbial community composition [55]. Three methods of whole soil metagenome analysis based on high-throughput DNA sequencing were assessed; shotgun, whole genome amplification and arbitrarily primed PCR. The metagenomic datasets underwent comprehensive statistical analysis using unconstrained and constrained approaches including CLUSTER analysis, NMDS and CAP ordination at all levels of both taxonomic and metabolic classification. The shotgun and WGA-based approaches generated highly similar metagenomic profiles for soil samples such that the soil samples could not be distinguished. An AP-PCR-based approach was shown to be the most powerful technique for obtaining site-specific metagenomic DNA profiles which were able

to successfully discriminate between similar soil samples taken from different locations.

The methods presented in this study show a significant step towards possible implementation of forensic soil discrimination using random whole metagenomics for investigation and evidence generation. By increasing the amount of samples analysed from each location and also by increasing the number of distinct geographical locations it will become possible to train algorithms that can then be used for comparison to unknown soil samples obtained as part of criminal investigations. The power of discrimination of these tools is proportional to the amount of samples taken and ultimately the unique metagenomic profile of the different locations. The investigation of temporal microbial variation would further strengthen any tool that is developed. As the sample sizes increase the tool will move from the model developed in this study to one that has sufficient power as a useful investigative tool and ultimately to a method that can be presented in court. The step to being a useful investigative tool for law enforcement can be made from the current study with increased repetition and geographic sampling. For presentation in a court of law the development of a sufficient sample size and distinct geographic profiles will need to be bolstered with a determination of the limitations of the method, including false positive and negative rates. This can be achieved via blind trials, mock case work and a period of casework hardening in order to achieve the levels require for acceptance.

## Supporting Information

**Figure S1 Rarefaction curves created in MG-RAST.** Rarefaction analysis was performed at the species level for each metagenomic protein-derived taxonomic profile based on randomly sub-samples datasets (49 000 reads per sample). The curves for all taxa include Bacteria, Archaea, Eukaryota, Viruses, unclassified and other sequences identified after metagenomic dataset annotation with M5NR database.
(TIF)

**Figure S2 Comparison of the soil protein-derived taxonomic profiles generated on full datasets at the class (A, B, C) and order (D, E, F) taxonomic resolution levels.** Bray-Curtis distance similarity matrix was calculated from the square-root transformed abundance of DNA fragments matching taxa to the M5NR database (E-value $<1\times10^{-5}$). The Bray-Curtis matrix was used for generating CLUSTER dendrogram, NMDS and CAP ordination plots. **CLUSTER analysis (A and D)**. Red dotted branches on the CLUSTER dendrogram indicate no significant difference between metagenomic profiles (supported by the SIMPROF analysis, $p<0.05$). **NMDS unconstrained ordination (B and E)**. The NMDS plot displays distances between samples. Data points that are closer to each other represent samples with highly similar metagenomic profiles. **CAP constrained ordination (C and F)**. CAP analysis tests for differences among the groups in multivariate space. The significance of group separation along the canonical axis is indicated by the value of the squared canonical correlation ($\delta_1^2$) and P-value ($P<0.05$). A contour line on the NMDS and CAP ordinations drawn round each of the cluster defines the superimposition of clusters from CLUSTER dendrogram at the selected level of similarity.
(TIF)

**Figure S3 Comparison of the soil rRNA profiles generated on full datasets at the phylum (A, B, C) and class (D, E, F) taxonomic resolution levels.** Bray-Curtis distance similarity matrix was calculated from the square-root transformed abundance of DNA fragments matching taxa in the M5RNA database (E-value $<1\times10^{-5}$). The Bray-Curtis matrix was used for generating CLUSTER dendrogram, NMDS and CAP ordination plots. **CLUSTER analysis (A and D)**. Red dotted branches on the CLUSTER dendrogram indicate no significant difference between metagenomic profiles (supported by the SIMPROF analysis, $p<0.05$). **NMDS unconstrained ordination (B and E)**. The NMDS plot displays distances between samples. Data points that are closer to each other represent samples with highly similar metagenomic profiles. **CAP constrained ordination (C and F)**. CAP analysis tests for differences among the groups in multivariate space. The significance of group separation along the canonical axis is indicated by the value of the squared canonical correlation ($\delta_1^2$) and P-value ($P<0.05$). A contour line on the NMDS and CAP ordinations drawn round each of the cluster defines the superimposition of clusters from CLUSTER dendrogram at the selected level of similarity.
(TIF)

**Figure S4 Comparison of the soil rRNA profiles generated on full datasets at the phylum (A, B, C) and class (D, E, F) taxonomic resolution levels.** Bray-Curtis distance similarity matrix was calculated from the square-root transformed abundance of DNA fragments matching taxa in the M5RNA database (E-value $<1\times10^{-5}$). The Bray-Curtis matrix was used for generating CLUSTER dendrogram, NMDS and CAP ordination plots. **CLUSTER analysis (A and D)**. Red dotted branches on the CLUSTER dendrogram indicate no significant difference between metagenomic profiles (supported by the SIMPROF analysis, $p<0.05$). **NMDS unconstrained ordination (B and E)**. The NMDS plot displays distances between samples. Data points that are closer to each other represent samples with highly similar metagenomic profiles. **CAP constrained ordination (C and F)**. CAP analysis tests for differences among the groups in multivariate space. The significance of group separation along the canonical axis is indicated by the value of the squared canonical correlation ($\delta_1^2$) and P-value ($P<0.05$). A contour line on the NMDS and CAP ordinations drawn round each of the cluster defines the superimposition of clusters from CLUSTER dendrogram at the selected level of similarity.
(TIF)

**Figure S5 Comparison of the soil rRNA profiles generated on full datasets at the order (A, B, C) and family (D, E, F) taxonomic resolution levels.** Bray-Curtis distance similarity matrix was calculated from the square-root transformed abundance of DNA fragments matching taxa in the M5RNA database (E-value $<1\times10^{-5}$). The Bray-Curtis matrix was used for generating CLUSTER dendrogram, NMDS and CAP ordination plots. **CLUSTER analysis (A and D)**. Red dotted branches on the CLUSTER dendrogram indicate no significant difference between metagenomic profiles (supported by the SIMPROF analysis, $p<0.05$). **NMDS unconstrained ordination (B and E)**. The NMDS plot displays distances between samples. Data points that are closer to each other represent samples with highly similar metagenomic profiles. **CAP constrained ordination (C and F)**. CAP analysis tests for differences among the groups in multivariate space. The significance of group separation along the canonical axis is indicated by the value of the squared canonical correlation ($\delta_1^2$) and P-value ($P<0.05$). A contour line on the NMDS and CAP ordinations drawn round each of the cluster defines the

superimposition of clusters from CLUSTER dendrogram at the selected level of similarity.
(TIF)

**Figure S6 Comparison of the soil rRNA profiles generated on full datasets at the genus (A, B, C) and species (D, E, F) taxonomic resolution levels.** Bray-Curtis distance similarity matrix was calculated from the square-root transformed abundance of DNA fragments matching taxa in the M5RNA database (E-value $<1\times10^{-5}$). The Bray-Curtis matrix was used for generating CLUSTER dendrogram, NMDS and CAP ordination plots. **CLUSTER analysis (A and D)**. Red dotted branches on the CLUSTER dendrogram indicate no significant difference between metagenomic profiles (supported by the SIMPROF analysis, $p<0.05$). **NMDS unconstrained ordination (B and E)**. The NMDS plot displays distances between samples. Data points that are closer to each other represent samples with highly similar metagenomic profiles. **CAP constrained ordination (C and F)**. CAP analysis tests for differences among the groups in multivariate space. The significance of group separation along the canonical axis is indicated by the value of the squared canonical correlation ($\delta_1^2$) and P-value ($P<0.05$). A contour line on the NMDS and CAP ordinations drawn round each of the cluster defines the superimposition of clusters from CLUSTER dendrogram at the selected level of similarity.
(TIF)

**Figure S7 Comparison of the soil metabolic profiles generated on full datasets at the subsystems level 2 (A, B, C) and level 3 (D, E, F) metabolic resolution levels.** Bray-Curtis distance similarity matrix was calculated from the square-root transformed abundance of DNA fragments matching taxa in the SEED database (E-value $<1\times10^{-5}$). The Bray-Curtis matrix was used for generating CLUSTER dendrogram, NMDS and CAP ordination plots. **CLUSTER analysis (A and D)**. Red dotted branches on the CLUSTER dendrogram indicate no significant difference between metagenomic profiles (supported by the SIMPROF analysis, $p<0.05$). **NMDS unconstrained ordination (B and E)**. The NMDS plot displays distances between samples. Data points that are closer to each other represent samples with highly similar metagenomic profiles. **CAP constrained ordination (C and F)**. CAP analysis tests for differences among the groups in multivariate space. The significance of group separation along the canonical axis is indicated by the value of the squared canonical correlation ($\delta_1^2$) and P-value ($P<0.05$). A contour line on the NMDS and CAP ordinations drawn round each of the cluster defines the superimposition of clusters from CLUSTER dendrogram at the selected level of similarity.
(TIF)

**Figure S8 Comparison of the soil protein-derived taxonomic profiles generated on randomly sub-sampled datasets at the phylum (A, B, C) and class (D, E, F) metabolic resolution levels.** Bray-Curtis distance similarity matrix was calculated from the square-root transformed abundance of DNA fragments matching taxa in the M5NR database (E-value $<1\times10^{-5}$). The Bray-Curtis matrix was used for generating CLUSTER dendrogram, NMDS and CAP ordination plots. **CLUSTER analysis (A and D)**. Red dotted branches on the CLUSTER dendrogram indicate no significant difference between metagenomic profiles (supported by the SIMPROF analysis, $p<0.05$). **NMDS unconstrained ordination (B and E)**. The NMDS plot displays distances between samples. Data points that are closer to each other represent samples with highly similar metagenomic profiles. **CAP constrained ordination (C and F)**. CAP analysis tests for differences among the groups in multivariate space. The significance of group separation along the canonical axis is indicated by the value of the squared canonical correlation ($\delta_1^2$) and P-value ($P<0.05$). A contour line on the NMDS and CAP ordinations drawn round each of the cluster defines the superimposition of clusters from CLUSTER dendrogram at the selected level of similarity.
(TIF)

**Figure S9 Comparison of the soil protein-derived taxonomic profiles generated on randomly sub-sampled datasets at the order (A, B, C) and family (D, E, F) taxonomic resolution levels.** Bray-Curtis distance similarity matrix was calculated from the square-root transformed abundance of DNA fragments matching taxa in the SEED database (E-value $<1\times10^{-5}$). The Bray-Curtis matrix was used for generating CLUSTER dendrogram, NMDS and CAP ordination plots. **CLUSTER analysis (A and D)**. Red dotted branches on the CLUSTER dendrogram indicate no significant difference between metagenomic profiles (supported by the SIMPROF analysis, $p<0.05$). **NMDS unconstrained ordination (B and E)**. The NMDS plot displays distances between samples. Data points that are closer to each other represent samples with highly similar metagenomic profiles. **CAP constrained ordination (C and F)**. CAP analysis tests for differences among the groups in multivariate space. The significance of group separation along the canonical axis is indicated by the value of the squared canonical correlation ($\delta_1^2$) and P-value ($P<0.05$). A contour line on the NMDS and CAP ordinations drawn round each of the cluster defines the superimposition of clusters from CLUSTER dendrogram at the selected level of similarity.
(TIF)

**Figure S10 Comparison of the soil protein-derived taxonomic profiles generated on randomly sub-sampled datasets at the genus (A, B, C) and species (D, E, F) taxonomic resolution levels.** Bray-Curtis distance similarity matrix was calculated from the square-root transformed abundance of DNA fragments matching taxa in the M5NR database (E-value $<1\times10^{-5}$). The Bray-Curtis matrix was used for generating CLUSTER dendrogram, NMDS and CAP ordination plots. **CLUSTER analysis (A and D)**. Red dotted branches on the CLUSTER dendrogram indicate no significant difference between metagenomic profiles (supported by the SIMPROF analysis, $p<0.05$). **NMDS unconstrained ordination (B and E)**. The NMDS plot displays distances between samples. Data points that are closer to each other represent samples with highly similar metagenomic profiles. **CAP constrained ordination (C and F)**. CAP analysis tests for differences among the groups in multivariate space. The significance of group separation along the canonical axis is indicated by the value of the squared canonical correlation ($\delta_1^2$) and P-value ($P<0.05$). A contour line on the NMDS and CAP ordinations drawn round each of the cluster defines the superimposition of clusters from CLUSTER dendrogram at the selected level of similarity.
(TIF)

**Figure S11 Comparison of the soil metabolic profiles generated on randomly sub-sampled datasets at the subsystems level 1 (A, B, C) and subsystems Level 2 (D, E, F) metabolic resolution levels.** Bray-Curtis distance similarity matrix was calculated from the square-root transformed abundance of DNA fragments matching taxa in the SEED database (E-value $<1\times10^{-5}$). The Bray-Curtis matrix was used for generating CLUSTER dendrogram, NMDS and CAP

ordination plots. **CLUSTER analysis (A and D)**. Red dotted branches on the CLUSTER dendrogram indicate no significant difference between metagenomic profiles (supported by the SIMPROF analysis, $p<0.05$). **NMDS unconstrained ordination (B and E)**. The NMDS plot displays distances between samples. Data points that are closer to each other represent samples with highly similar metagenomic profiles. **CAP constrained ordination (C and F)**. CAP analysis tests for differences among the groups in multivariate space. The significance of group separation along the canonical axis is indicated by the value of the squared canonical correlation ($\delta_1^2$) and P-value ($P<0.05$)… A contour line on the NMDS and CAP ordinations drawn round each of the cluster defines the superimposition of clusters from CLUSTER dendrogram at the selected level of similarity.
(TIF)

**Figure S12 Comparison of the soil metabolic profiles generated on randomly sub-sampled datasets at the subsystems level 3 (A, B, C) and subsystems functions (D, E, F) metabolic resolution levels.** Bray-Curtis distance similarity matrix was calculated from the square-root transformed abundance of DNA fragments matching taxa in the SEED database (E-value $<1\times10^{-5}$). The Bray-Curtis matrix was used for generating CLUSTER dendrogram, NMDS and CAP ordination plots. **CLUSTER analysis (A and D)**. Red dotted branches on the CLUSTER dendrogram indicate no significant difference between metagenomic profiles (supported by the SIMPROF analysis, $p<0.05$). **NMDS unconstrained ordination (B and E)**. The NMDS plot displays distances between samples. Data points that are closer to each other represent samples with highly similar metagenomic profiles. **CAP constrained ordination (C and F)**. CAP analysis tests for differences among the groups in multivariate space. The significance of group separation along the canonical axis is indicated by the value of the squared canonical correlation ($\delta_1^2$) and P-value ($P<0.05$). A contour line on the NMDS and CAP ordinations drawn round each of the cluster defines the superimposition of clusters from CLUSTER dendrogram at the selected level of similarity.
(TIF)

**Table S1 Summary of soil metagenomic samples.** All metagenomes are publically available on the MG-RAST server (http://metagenomics.anl.gov/).
(PDF)

**Table S2 Protein-derived taxonomic composition of the soil microbial communities.** Relative abundances of major taxa (phylum level) derived from taxonomic assignment of protein gene fragments matched to M5NR database.
(PDF)

**Table S3 Taxonomic composition of the soil microbial communities based on rRNA gene fragments classification.** Relative abundances of major taxa (phylum level) derived from taxonomic assignment of rRNA gene fragments matched to M5RNA database.
(PDF)

**Table S4 Metabolic composition of the soil microbial communities.** Relative abundances of major metabolic features (level 1) derived from annotation of protein gene fragments to SEED Subsystems database.
(PDF)

**Table S5 Results of CAP model cross-validation of soil protein-derived taxonomic profiles discrimination generated from sub-sampled sequencing datasets.**
(PDF)

**Table S6 Results of CAP model cross-validation of soil metabolic profiles discrimination generated from sub-sampled sequencing datasets.**
(PDF)

**Table S7 Results of CAP model cross-validation of soil rRNA taxonomic profiles discrimination generated from full sequencing datasets.**
(PDF)

## Acknowledgments

We would like to thank Dr. Paul Gooding (AGRF, Adelaide) for HTS sequencing and Dr. Sophie Leterme (Flinders University) for helpful assistance in the statistical analysis of the metagenomic data.

## Author Contributions

Conceived and designed the experiments: ASK LB AL. Performed the experiments: ASK. Analyzed the data: ASK RJS. Contributed reagents/materials/analysis tools: LB AL. Contributed to the writing of the manuscript: ASK RJS DA AL.

## References

1. Fitzpatrick R (2009) Soil: Forensic Analysis. In: Jamieson A, Moenssens A, editors. Wiley Encyclopedia of Forensic Science. The Atrium, Southern Gate, Chichester, West Sussex, United Kingdom: John Wiley&Sons Ltd. 2377 2388.
2. Dawson LA, Hillier S (2010) Measurement of soil characteristics for forensic applications. Surf and Interface Anal 42: 363 377.
3. Cox RJ, Peterson HL, Young J, Cusik C, Espinoza EO (2000) The forensic analysis of soil organic by FTIR. Forensic Sci Int 108: 107 116.
4. Ruffell A, Wiltshire P (2004) Conjunctive use of quantitative and qualitative X-ray diffraction analysis of soils and rocks for forensic analysis. Forensic Sci Int 145: 13 23.
5. Moreno LI, Mills DK, Entry J, Sautter RT, Mathee K (2006) Microbial metagenome profiling using amplicon length heterogeneity-polymerase chain reaction proves more effective than elemental analysis in discriminating soil specimens. J Forensic Sci 51: 1315 1322.
6. Arroyo L, Trejos T, Hosick T, Machemer S, Almirall JR, et al. (2010) Analysis of Soils and Sediments by Laser Ablation Inductively Coupled Plasma Mass Spectrometry (LA-ICP-MS): An Innovative Tool for Environmental Forensics. Environ Forensics 11: 315 327.
7. Macdonald CA, Ang R, Cordiner SJ, Horswell J (2011) Discrimination of soils at regional and local levels using bacterial and fungal T-RFLP profiling. J Forensic Sci 56: 61 69.
8. Horswell J, Cordiner SJ, Maas EW, Martin TM, Sutherland KBW, et al. (2002) Forensic comparison of soils by bacterial community DNA profiling. J Forensic Sci 47: 350 353.
9. Sensabaugh GF (2009) Microbial Community Profiling for the Characterisation of Soil Evidence: Forensic Considerations. In: Ritz K, Dawson L, Miller D, editors. Criminal and Environmental Soil Forensics. Springer. 49 60.
10. Lenz EJ, Foran DR (2010) Bacterial profiling of soil using genus-specific markers and multidimensional scaling. J Forensic Sci 55: 1437 1442.
11. Pasternak Z, Al-Ashhab A, Gatica J (2012) Optimization of molecular methods and statistical procedures for forensic fingerprinting of microbial soil communities. Int Res J Microbiol 3: 363 372.
12. Concheri G, Bertoldi D, Polone E, Otto S, Larcher R, et al. (2011) Chemical elemental distribution and soil DNA fingerprints provide the critical evidence in murder case investigation. PloS One 6: e20222.
13. Rincon-Florez V, Carvalhais L, Schenk P (2013) Culture-Independent Molecular Tools for Soil and Rhizosphere Microbiology. Diversity 5: 581 612.
14. Shokralla S, Spall JL, Gibson JF, Hajibabaei M (2012) Next-generation sequencing technologies for environmental DNA research. Mol Ecol 21: 1794 1805.
15. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, et al. (2012) Performance comparison of benchtop high-throughput sequencing platforms. Nat Biotechnol 30: 434 439.

16. Logares R, Haverkamp THA, Kumar S, Lanzén A, Nederbragt AJ, et al. (2012) Environmental microbiology through the lens of high-throughput DNA sequencing: synopsis of current platforms and bioinformatics approaches. J Microbiol Methods 91: 106 113.
17. Fierer N, Leff JW, Adams BJ, Nielsen UN, Bates ST, et al. (2012) Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. Proc Natl Acad Sci USA 109: 21390 21395.
18. Xu Z, Hansen MA, Hansen LH, Jacquiod S, Sørensen SJ (2014) Bioinformatic approaches reveal metagenomic characterization of soil microbial community. PloS One 9: e93445.
19. Wang J, McLenachan P A, Biggs PJ, Winder LH, Schoenfeld BIK, et al. (2013) Environmental bio-monitoring with high-throughput sequencing. Brief Bioinform 14: 575 588.
20. Fierer N, Lauber CL, Ramirez KS, Zaneveld J, Bradford M A, et al. (2012) Comparative metagenomic, phylogenetic and physiological analyses of soil microbial communities across nitrogen gradients. ISME J 6: 1007 1017.
21. Suenaga H (2012) Targeted metagenomics: a high-resolution metagenomics approach for specific gene clusters in complex microbial communities. Environ Microbiol 14: 13 22.
22. Delmont TO, Prestat E, Keegan KP, Faubladier M, Robe P, et al. (2012) Structure, fluctuation and magnitude of a natural grassland soil metagenome. ISME J 6: 1677 1687.
23. Prakash T, Taylor TD (2012) Functional assignment of metagenomic data: challenges and applications. Brief Bioinform 13: 711 727.
24. Fuhrman JA (2012) Metagenomics and its connection to microbial community organization. F1000 Biol Rep 4: 15.
25. Howe AC, Jansson JK, Malfatti SA, Tringe SG, Tiedje JM, et al. (2014) Tackling soil diversity with the assembly of large, complex metagenomes. Proc Natl Acad Sci USA 111: 4904 4909.
26. Young JM, Rawlence NJ, Weyrich LS, Cooper A (2014) Limitations and recommendations for successful DNA extraction from forensic soil samples: A review. Sci Justice 54: 238 244.
27. Sagar K, Singh SP, Goutam KK, Konwar BK (2014) Assessment of five soil DNA extraction methods and a rapid laboratory-developed method for quality soil DNA extraction for 16S rDNA-based amplification and library construction. J Microbiol Methods 97: 68 73.
28. Terrat S, Christen R, Dequiedt S, Lelièvre M, Nowak V, et al. (2012) Molecular biomass and MetaTaxogenomic assessment of soil microbial communities as influenced by soil DNA extraction procedure. Microbiol Biotech 5: 135 141.
29. Binga EK, Lasken RS, Neufeld JD (2008) Something from (almost) nothing: the impact of multiple displacement amplification on microbial ecology. ISME J 2: 233 241.
30. Mao D-P, Zhou Q, Chen C-Y, Quan Z-X (2012) Coverage evaluation of universal bacterial primers using the metagenomic datasets. BMC Microbiol 12: 66.
31. Lewis C, Bilkhu S (2011) Identification of Fungal DNA Barcode Targets and PCR Primers Based on Pfam Protein Families and Taxonomic Hierarchy. Open Applied Informatics: 30 44.
32. Bates ST, Clemente JC, Flores GE, Walters WA, Parfrey LW, et al. (2013) Global biogeography of highly diverse protistan communities in soil. ISME J 7: 652 659.
33. Lauber CL, Ramirez KS, Aanderud Z, Lennon J, Fierer N (2013) Temporal variability in soil microbial communities across land-use types. ISME J 7: 1641 1650.
34. Morrisson A, McColl S, Dawson L (2009) Characterisation and Discrimination of Urban Soils: Preliminary Results from the Soil Forensics University Network. Soil Forensics. In: Ritz K, Dawson L, Miller D, editors. Criminal and Environmental Soil Forensics. Springer. 75 86.
35. Khodakova AS, Burgoyne L, Abarno D, Linacre A (2013) Forensic analysis of soils using single arbitrarily primed amplification and high throughput sequencing. Forensic Sci Int Genet Suppl Ser 4: e39 e40.
36. Welsh J, McClelland M (1990) Fingerprinting genomes using PCR with arbitrary primers. Nucleic Acids Res 18: 7213 7218.
37. Caetano-Anolles G (1993) Amplifying DNA with arbitrary oligonucleotide primers. Genome Res 3: 85 94.
38. Dabrowski W, Czekajło-Kołodziej U, Medrala D, Giedrys-Kalemba S (2003) Optimisation of AP-PCR fingerprinting discriminatory power for clinical isolates of Pseudomonas aeruginosa. FEMS Microbiol Lett 218: 51 57.
39. Roy S, Biswas D, Vijayachari P, Sugunan A P, Sehgal SC (2004) A 22-mer primer enhances discriminatory power of AP-PCR fingerprinting technique in characterization of leptospires. Trop Med Int Health9: 1203 1209.
40. Franklin RB, Taylor DR, Mills AL (1999) Characterization of microbial communities using randomly amplified polymorphic DNA (RAPD). J Microbiol Methods 35: 225 235.
41. Waters JM, Eariss G, Yeadon PJ, Kirkbride KP, Burgoyne LA, et al. (2012) Arbitrary single primer amplification of trace DNA substrates yields sequence content profiles that are discriminatory and reproducible. Electrophoresis 33: 492 498.
42. Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, et al. (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinform 9: 386.
43. Clarke K, Gorley R (2006) PRIMER V6: User Manual/Tutorial.
44. Clarke K (1993) Non-parametric multivariate analyses of changes in community structure. Aust J Ecol: 117 143.
45. Anderson MJ, Gorley RN CK (2008) PERMANOVA+ for PRIMER: Guide to Software and Statistical Methods.
46. Anderson M, Willis T (2003) Canonical analysis of principal coordinates: a useful method of constrained ordination for ecology. Ecology 84: 511 525.
47. Gilbert JA, Field D, Swift P, Thomas S, Cummings D, et al. (2010) The taxonomic and functional diversity of microbes at a temperate coastal site: a "multi-omic" study of seasonal and diel temporal variation. PloS One 5: e15545.
48. Shange R, Haugabrooks E, Ankumah R, Ibekwe A, Smith R, et al. (2013) Assessing the Diversity and Composition of Bacterial Communities across a Wetland, Transition, Upland Gradient in Macon County Alabama. Diversity 5: 461 478.
49. Uroz S, Ioannidis P, Lengelle J, Cébron A, Morin E, et al. (2013) Functional assays and metagenomic analyses reveals differences between the microbial communities inhabiting the soil horizons of a Norway spruce plantation. PloS One 8: e55929.
50. Atienzar F, Jha A (2006) The random amplified polymorphic DNA (RAPD) assay and related techniques applied to genotoxicity and carcinogenesis studies: a critical review. Mutat Res 613: 76 102.
51. Jeffries TC, Seymour JR, Gilbert JA, Dinsdale E A, Newton K, et al. (2011) Substrate type determines metagenomic profiles from diverse chemical habitats. PloS One 6: e25173.
52. Håvelsrud OE, Haverkamp THA, Kristensen T, Jakobsen KS, Rike AG (2012) Metagenomic and geochemical characterization of pockmarked sediments overlaying the Troll petroleum reservoir in the North Sea. BMC Microbiol 12: 203.
53. Smith RJ, Jeffries TC, Adetutu EM, Fairweather PG, Mitchell JG (2013) Determining the metabolic footprints of hydrocarbon degradation using multivariate analysis. PloS One 8: e81940.
54. Zarraonaindia I, Smith DP, Gilbert JA (2013) Beyond the genome: community-level analysis of the microbial world. Biol Philos 28: 261 282.
55. Delmont TO, Simonet P, Vogel TM (2012) Describing microbial communities and performing global comparisons in the 'omic era. ISME J 6: 1625 1628.

# Appendix D. Additional publications in peer-reviewed journals

- D.A. Khodakov, **A.S. Khodakova**, D. Huang, A. Linacre, A.V. Ellis. Protected DNA strand displacement for enhanced single nucleotide discrimination in double-stranded DNA. **Scientific Reports, 2015**, Accepted, In Press.

- D.A. Khodakov, **A.S. Khodakova**, A. Linacre, A.V. Ellis. Sequence selective capture, release and analysis of DNA using a magnetic microbead-assisted toehold-mediated DNA strand displacement reaction. **Analyst, 2014**, 139(14): 3548-3551.

Analyst

COMMUNICATION

View Article Online
View Journal



# Sequence selective capture, release and analysis of DNA using a magnetic microbead-assisted toehold-mediated DNA strand displacement reaction†

Cite this: DOI: 10.1039/c4an00694a

Received 17th April 2014
Accepted 15th May 2014

DOI: 10.1039/c4an00694a

www.rsc.org/analyst

Dmitriy A. Khodakov,*[a] Anastasia S. Khodakova,[b] Adrian Linacre[b] and Amanda V. Ellis*[a]

This paper reports on the modification of magnetic beads with oligonucleotide capture probes with a specially designed pendant toehold (overhang) aimed specifically to capture double-stranded PCR products. After capture, the PCR products were selectively released from the magnetic beads by means of a toehold-mediated strand displacement reaction using short artificial oligonucleotide triggers and analysed using capillary electrophoresis. The approach was successfully shown on two genes widely used in human DNA genotyping, namely human c-fms (macrophage colony-stimulating factor) proto-oncogene for the CSF-1 receptor (CSF1PO) and amelogenin.

Selective capture of DNA on two- and three-dimensional surfaces is the underlying principle of a vast majority of nucleic acid molecular analytical techniques such as high density microarrays, nucleic acid biosensors, and personalised point-of-care micro-devices. These approaches use surfaces modified with immobilised capture probes for the hybridisation of nucleic acids, which is regarded as the final point of analysis. Patterned surface modification allows for the localisation of specific DNA probes at different regions, which in turn allows for highly efficient selection, separation and enrichment of complementary DNA molecules (targets) from the initial crude mixture.

More recently however, in light of advances in the field of gene therapy,[1] there has been a great deal of interest in the development of methods for controlled (stimuli-responsive) release of nucleic acids from both 2- and 3-dimensional surfaces. It is envisaged that such systems will facilitate the advent of smart stimuli-responsive gene delivery vehicles. To date, several different methods exploiting such external stimuli have been reported, employing the use of temperature,[2] DC electric field,[3] light,[4] and low molecular weight organic molecules.[5] Temperature controlled DNA release relies on thermal denaturation of DNA double helices trapped on a surface *via* immobilised ligands (capture probes), such as another DNA molecule[6] or avidin–biotin complexes.[2b] This approach possesses low specificity and requires either harsh dehybridisation conditions or fine tuning of the system to be able to operate specifically at physiological temperatures. The application of a DC electric field has also shown potential for the fast and specific dehybridisation of 3′-mismatched DNA duplexes from electro-conductive surfaces.[3c]

Several publications[4d,e] describe the selective release of multiple DNA molecules. These systems are mainly based on the use of gold nanoparticles with varying morphologies, and thus surface plasmon absorbances. Irradiation of gold nanoparticle mixtures, loaded with covalently captured DNA duplexes, at their specific plasmon resonance frequency causes preferential heating of some particles but not others. This consequently leads to melting of DNA duplexes and their subsequent release from the nanoparticle. In order to have specificity of DNA release, there must be no overlapping of the surface plasmon resonances and this is far from trivial.

The discovery of DNA strand-displacement reactions, prompted by the formation of a toehold structure (for references see ref. 7), has opened a new era in programmable synthetic biology.[8] One of the most attractive features of this reaction is that short artificial DNA strands allow for a specific trigger of different downstream scenarios which include both nucleic acid and protein cascades.[9] Applying these systems to solid phase supports has recently gained much interest.[10] For example, 2-dimensional microarrays operated by a strand-displacement reaction were shown to be successful at DNA length polymorphism measurements (short tandem repeat analysis, STR).[10e] Picuri *et al.*[9a] have applied the DNA displacement reaction on sepharose microbeads to translate several specific biologically relevant DNA and RNA molecules (HIV,

[a]*Flinders Centre for Nanoscale Science and Technology, Flinders University, GPO Box 2100, Adelaide, SA, 5001 Australia. E-mail: dmitriy.khodakov@flinders.edu.au; amanda.ellis@flinders.edu.au*

[b]*School of Biological Sciences, Flinders University, GPO Box 2100, Adelaide, SA, 5001 Australia*

† Electronic supplementary information (ESI) available: Additional experimental details and results. See DOI: 10.1039/c4an00694a

HCV and smallpox) into an unique unspecific DNA output triggering diagnostic assay. Probst *et al.*[10c] covalently immobilised antibody encoded double stranded (ds)-DNA with pendant toeholds onto magnetic beads. After binding with specific antigens, the antibody–antigen complexes were released from the bead surface upon addition of specific displacement probes.

Here we report an approach for the capture of multiple ds-DNA molecules on a surface of magnetic beads followed by their sequence selective release using a toehold-mediated strand-displacement reaction. To achieve this we modified magnetic beads with single-stranded oligonucleotides (ss-ODNs) (capture probes). These capture probes were specially designed to capture specific ds-PCR amplification products, tagged with a PEGylated (polyethylene glycol) ss-ODN, of forensically relevant human c-fms (macrophage colony-stimulating factor) proto-oncogene for the CSF-1 receptor (CSF1PO) and amelogenin (AMEL) genes. STR analysis using capillary electrophoresis (CE) is the main form of human identification used in forensics and considered as the 'gold standard' of forensic science.[11a] STRs have been used in forensic science since 1994 including the adoption of CSF1PO,[11b] which was developed originally by the Promega Corporation. This locus is an example of a simple repeat motif $(AGAT)_n$ and was initially incorporated into the US Combined DNA Index System (CODIS) loci. CSF1PO continues to be one of the loci used in commercial kits by the forensic community worldwide. Amelogenin is one of the few genes with homologues on the X and Y chromosomes[11b] and was adopted in 1995 into STR multiplexes and used ever since. The X chromosome version has a 6 bp deletion compared to the Y homologue allowing this size variation to be used to determine the gender of the person from whom the DNA profile was generated.

Each capture probe had a main hybridisation sequence (Fig. 1, domain *a*) with an extra 8 nucleotide long ss-ODN (toehold) attached at the 5′ termini of the main hybridisation sequence (Fig. 1, domain *b*). The entire capture probe was immobilised covalently to *N*-hydroxysuccinimide activated carboxy-terminated magnetic beads (1 μm, Bioclone, USA) using amide coupling through the 3′ end of the amine-modified capture probes.

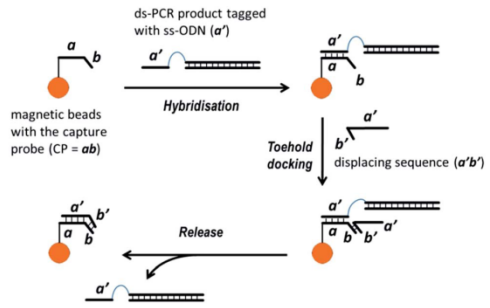



Fig. 1 Schematic of the capture of the ds-PCR products tagged with ss-ODN with following sequences selective release *via* toehold-mediated DNA strand displacement reaction. Domains *a* and *a′*, and *b* and *b′* are self-complementary.

The synthesis of the ds-PCR products tagged with a PEGylated ss-ODN (Fig. 1, domain *a′*) has been previously described by our group.[12] In brief, the forward primer for amplification of the target genes consisted of the primer's main sequence (not shown), an additional shorter sequence (Fig 1, domain *a′*) (for hybridisation to the capture probe (Fig. 1, domain *a*)) and a PEG spacer linking at the 5′-termini of the primer's main sequence and the 3′-termini of the hybridisation sequence. At the completion of the PCR all the products of amplification contained a short tail of PEGylated ss-ODN sequences (attached to the amplified ds-PCR product) which could be hybridised (captured) to the surface immobilised capture probes. Importantly, all unreacted PEGylated primers form a self-complementary intramolecular hairpin structure that prevents hybridisation of any unreacted primers with the surface immobilised capture probes.

Fig. 1 shows the scheme of the capture and sequence selective release of ds-PCR products with PEGylated ss-ODN on the magnetic beads. The ds-PCR product tagged with PEGylated ss-ODN tail (Fig. 1 domain *a′*) first hybridises with the capture probe covalently immobilised on the surface of the magnetic beads. In particular, during this hybridisation step the ss-ODN (domain *a′*) specifically interacts with domain *a* of the capture probe thus forming a perfectly matched DNA duplex *aa′* and at the same time leaving domain *b* (toehold) free. After several washings of the magnetic beads, now bearing the ds-PCR product, a specific displacing sequence consisting of domains *a′b′* is added. The domains *b* and *b′* then hybridise to each other forming a toehold structure (Fig. 1 complex *bb′*) which then promotes the strand displacement of the domain *a′* from the initial hybridisation duplex, resulting in sequence selective release of the ds-PCR product from the magnetic bead surface back into solution. The results of the displacement reactions and the length analysis of the released ds-PCR products were then confirmed using CE.

In order to evaluate the feasibility of the approach for a one-pot sequence selective release of multiple hybridised ds-DNA molecules, we applied this technique for the capture, sequential release, isolation and analysis of two ds-PCR products CSF1PO and AMEL genes. Gene-specific capture probes with corresponding displacing sequences (Table S1, ESI†) were designed to capture and release ss-ODN tagged ds-PCR products obtained after amplification with specific modified primers (as described above) of the CSF1PO and AMEL loci in human genomic DNA. Reverse primers for both the CSF1PO and AMEL genes were labelled with a fluorescent dye (fluorescein) at their 5′ termini in order to be observed during CE analysis. Multiplex PCR amplification with the designated primers (Table S1, ESI†) was carried out using 5 ng of male genomic DNA isolated from the author's own blood.

In order to capture the CSF1PO and AMEL ds-PCR amplification products (generated from male genomic DNA) with PEGylated ss-ODN the PCR mixture and hybridisation buffer (final concentration of guanidine thiocyanate (1 M), HEPES (50 mM), pH 7.5, and EDTA (5 mM)) were directly applied to a

stoichiometric mixture of the magnetic beads with immobilised CSF1PO and AMEL capture probes. Significantly, this negates laborious and time consuming PCR clean-up procedures. After hybridisation (Fig. 1), the magnetic beads were washed multiple times by removing them from the hybridisation solution with a magnet and washing with a washing buffer (6.5× SSPE buffer pH 7.4, 0.01% Tween 20) and water to remove any non-hybridised ds-PCR products on the bead surface.

Both the extent and stability of hybridisation were then determined by CE analysis. Fig. 2A shows the electropherogram of the sequences released after 18 h (overnight) incubation of the magnetic bead mixture in a 1× TEM (10 mM Tris–HCl pH 8.0, 1 mM EDTA, 12.5 mM $MgCl_2$) displacement buffer with no ss-ODN displacing sequence. Clearly, there has been no release of the ds-PCR products from the magnetic beads indicating that no toehold strand displacement reaction has taken place (Fig. 2A).

Following these results the first round of sequence selective release of the hybridised ds-PCR products was carried out. After washing, the magnetic bead mixture was resuspended in a solution at 30 °C that consisted of a 1× TEM displacement buffer and a displacing sequence (1 μM) fully complementary to the CSF1PO capture probe (Table S1, ESI†). During this stage domain $b'$ of the CSF1PO displacing sequence binds with the domain $b$ of the CSF1PO capture probe only (Fig. 1), but not the AMEL capture probe. Subsequently, this triggers the release of the CSF1PO ds-PCR product through a 3-way branch migration mechanism. An aliquot was taken from this reaction mixture after 0, 2, 4, 6 and 18 h (overnight) of incubation. However, there was only a noticeable level of ds-PCR product released *via* CE analysis after 18 h (overnight) of incubation (for 0, 2, 4 and 6 h of displacement data not shown). Slow release kinetics are likely due to steric hindrance and electrostatic repulsions caused by the long ds-PCR products, ranging from 100 to 344 bp, hybridised on the magnetic bead surface.

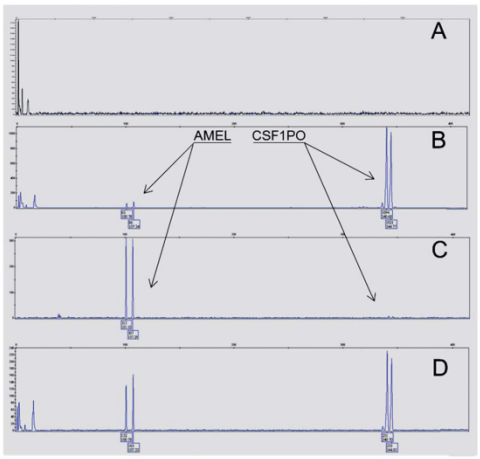



Fig. 2 CE electropherograms of the sequence selective release of ds-PCR products (generated from male DNA) dehybridised from magnetic beads using a toehold strand displacement reaction in the presence of (A) no displacing sequence, (B) a CSF1PO displacing sequence, (C) an AMEL displacing sequence and (D) both CSF1PO and AMEL displacing sequences.

Fig. 2B shows the CE electropherogram of the ds-PCR products present in solution after a CSF1PO displacing sequence was added to the magnetic bead mixture (18 h of incubation). It can clearly be seen that after incubation, two peaks with intensities of 1094 and 1021 relative fluorescence units (RFU) of the CSF1PO ds-PCR products were observed (Fig. 2B). The two peaks correspond to two different CSF1PO ds-PCR products (340 and 344 bp) amplified from two alleles of the CSF1PO locus. The length of these amplification products was in full accordance with the values determined by the PowerPlex 21 STR genotyping system (Promega, USA) (Fig. S1, ESI†). Unspecific AMEL release was observed at <85 RFU (<8.5% of the CSF1PO intensity value) (Fig. 2B).

The capability of the system to sequentially release a second hybridised ds-PCR amplification product was then evaluated. The same magnetic bead mixture (after CSF1PO ds-PCR products had been released) was rinsed thoroughly with 1× TEM buffer and suspended in an AMEL displacing solution consisting of 1× TEM buffer and AMEL displacing sequence (1 μM) (Table S1, ESI†). CE analysis was then performed on an aliquot of the solution after 2, 4, 6 and 18 h incubation of the magnetic beads. Surprisingly, the detectable level of the AMEL ds-PCR product release was achieved after 2 h of incubation. Fig. 2C shows the CE electropherogram of the displaced AMEL ds-PCR product with peak intensities of 313 and 307 RFU. In this particular case these two peaks are due to the amplification of the AMEL loci from the X and Y chromosomes (XY – male DNA genotype), respectively, previously confirmed by the PowerPlex 21 STR genotyping system (Fig. S1, ESI†). Unspecific displacement of the CSF1PO amplification products was <10 RFU (<3% of the AMEL intensity value).

It was believed that the difference in the displacement times (rates) between the first and the second rounds of the consecutive release were related to the washing step of the initial magnetic bead mixture with the washing buffer containing Tween-20 surfactant. To test this, the initial mixture of magnetic beads (with both PCR products hybridised) was washed with 1× TEM buffer only (no Tween-20 surfactant). The first round of the release with CSF1PO displacement sequence was then repeated. The CE analysis performed on aliquots taken in the same timeframes (2, 4, 6 and 18 h) showed the successful displacement after 6 h of incubation (*c.f.* 18 h previously) with intensities of the CSF1PO peaks of 216 and 198 RFU (Fig. S3, ESI†). The observed increase in the release rate is likely explained by the "shielding" effect of Tween-20 physically adsorbed on the magnetic bead surface. This may hinder the efficient formation of the toehold structure and the subsequent branch migration displacement. The reaction timeframes are not surprising, since similar kinetic release (2–72 h) has been observed by Baker *et al.* who used 15 nucleotide long duplexes immobilised on 5 μm carboxylated latex particles.[10a,b]

Our system was also capable of the simultaneous release of both CSF1PO and AMEL ds-PCR products from the magnetic beads. This was achieved by adding two displacing sequences at once (CSF1PO and AMEL). Fig. 2D shows the CE electropherogram of the aliquot taken after 6 h incubation of the magnetic bead mixture (with both PCR products hybridised) in 1× TEM buffer containing SCF1PO and AMEL displacing sequences (1 μM, each). As expected, two peaks for each of the PCR products were observed with intensities of 231 and 210 RFU for CSF1PO and 132 and 163 RFU for AMEL. The length of the released PCR products also fits the data shown above (Fig. 2B and C).

A final series of experiments showed the successful release of the surface captured ds-PCR products generated from the author's female genomic DNA (Fig. S4, ESI†). In this case the AMEL ds-PCR product was selectively released first by incubation of the hybridised beads in the AMEL displacing solution. CE analysis of the released product showed only one peak confirming the XX genotype of the female DNA sample (Fig. S4B†). Subsequent incubation of the beads in the CSF1PO displacing solution resulted in the release of the CSF1PO ds-PCR product which was then also analysed by CE. The single peak observed (Fig. S4C, ESI†) fits the CSF1PO genetic profile previously identified by the PowerPlex 21 STR genotyping system (Fig S2, ESI†). Simultaneous release of both PCR products was also shown to be successful (Fig. S4C, ESI†).

## Conclusions

In summary, we have developed a simple DNA sequence trigger approach to specifically release surface hybridised ds-PCR amplification products. The approach is based on the specific reaction of toehold-mediated DNA strand-displacement. The selective release of one out of two surface hybridised ds-PCR amplification products was manipulated by simple addition of a specific small ss-ODN which played the role of the trigger. The release of the second hybridised ds-PCR product was achieved simply by adding another specific ss-ODN trigger. The release was carried out under mild conditions (low salt Tris–EDTA buffer pH 8) at 30 °C showing that the approach could be readily implemented in *in vivo* experiments. Isolation and analysis of the released ds-PCR products showed the integrity of the DNA molecules. The method shows great promise for a broad range of other consecutive scenarios such as the development of highly specific stimuli-responsive molecular cargo vehicles operated by DNA or RNA.

## Notes and references

1 (*a*) S. L. Ginn, I. E. Alexander, M. L. Edelstein, M. R. Abedi and J. Wixon, *J. Gene Med.*, 2013, **15**, 65; (*b*) E.-Y. Kim, R. Schulz, P. Swantek, K. Kunstman, M. H. Malim and S. M. Wolinsky, *Gene Ther.*, 2012, **19**, 347353; (*c*) A. G. Bader, D. Brown, J. Stoudemire and P. Lammers, *Gene Ther.*, 2011, **18**, 1121.

2 (*a*) A. Ohsugi, H. Furukawa, A. Kakugo, Y. Osada and J. P. Gong, *Macromol. Rapid Commun.*, 2006, **27**, 1242; (*b*) S. H. Yeung, P. Liu, N. Del Bueno, S. A. Greenspoon and R. A. Mathies, *Anal. Chem.*, 2009, **81**, 210; (*c*) N. Thaitrong, P. Liu, T. Briese, W. I. Lipkin, T. N. Chiesl, Y. Higa and R. A. Mathies, *Anal. Chem.*, 2010, **82**, 10102.

3 (*a*) C. Gautier, C. Cougnon, J.-F. Pilard, N. Casse and B. Chénais, *Anal. Chem.*, 2007, **79**, 7920; (*b*) J. Wang, G. Rivas, M. Jiang and X. Zhang, *Langmuir*, 1999, **15**, 6541; (*c*) I. Y. Wong and N. A. Melosh, *Nano Lett.*, 2009, **9**, 3521.

4 (*a*) A. Barhoumi, R. Huschka, R. Bardhan, M. W. Knight and N. J. Halas, *Chem. Phys. Lett.*, 2009, **482**, 171; (*b*) F. L. Callari, S. Petralia, S. Conoci and S. Sortino, *New J. Chem.*, 2008, **32**, 1899; (*c*) W. Fischer, M. A. Quadir, A. Barnard, D. K. Smith and R. Haag, *Macromol. Biosci.*, 2011, **11**, 1736; (*d*) A. Wijaya, S. B. Schaffer, I. G. Pallares and K. Hamad-Schifferli, *ACS Nano*, 2009, **3**, 80; (*e*) R. Huschka, J. Zuloaga, M. W. Knight, L. V. Brown, P. Nordlander and N. J. Halas, *J. Am. Chem. Soc.*, 2011, **133**, 12247; (*f*) H. de Puig, A. Cifuentes Rius, D. Flemister, S. H. Baxamusa and K. Hamad-Schifferli, *PLoS One*, 2013, **8**, e68511.

5 (*a*) S. L. Ng, G. K. Such, A. P. R. Johnston, G. Antequera-García and F. Caruso, *Biomaterials*, 2011, **32**, 6277; (*b*) E. J. Moore, M. Curtin, J. Ionita, A. R. Maguire, G. Ceccone and P. Galvin, *Anal. Chem.*, 2007, **79**, 2050.

6 A. Bromberg, E. C. Jensen, J. Kim, Y. K. Jung and R. A. Mathies, *Anal. Chem.*, 2012, **84**, 963–970.

7 (*a*) D. Y. Zhang and G. Seelig, *Nat. Chem.*, 2011, **3**, 103; (*b*) N. Srinivas, T. E. Ouldridge, P. Sulc, J. M. Schaeffer, B. Yurke, A. A. Louis, J. P. K. Doye and E. Winfree, *Nucleic Acids Res.*, 2013, **41**, 10641.

8 (*a*) A. Prokup, J. Hemphill and A. Deiters, *J. Am. Chem. Soc.*, 2012, **134**, 3810; (*b*) I. K. Astakhova, K. Pasternak, M. A. Campbell, P. Gupta and J. Wengel, *J. Am. Chem. Soc.*, 2013, **135**, 2423; (*c*) J. Shin and N. Pierce, *Nano Lett.*, 2004, **4**, 905.

9 (*a*) J. M. Picuri, B. M. Frezza and M. R. Ghadiri, *J. Am. Chem. Soc.*, 2009, **131**, 9368; (*b*) X. Qi, C. Lu, X. Liu, S. Shimron, H. Yang and I. Willner, *Nano Lett.*, 2013, **13**, 4920.

10 (*a*) B. A. Baker, G. Mahmoudabadi and V. T. Milam, *Colloids Surf., B*, 2013, **102**, 884; (*b*) B. A. Baker and V. T. Milam, *Nucleic Acids Res.*, 2011, **39**, e99; (*c*) C. E. Probst, P. Zrazhevskiy and X. Gao, *J. Am. Chem. Soc.*, 2011, **133**, 17126; (*d*) H. Subramanian, B. Chakraborty, R. Sha and N. C. Seeman, *Nano Lett.*, 2011, **11**, 910; (*e*) N. Pourmand, S. Caramuta, A. Villablanca, S. Mori, M. Karhanek, S. X. Wang and R. W. Davis, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 6146; (*f*) B. Frezza, S. Cockroft and M. Ghadiri, *J. Am. Chem. Soc.*, 2007, **129**, 14875.

11 (*a*) J. Butler, *Advanced topics in forensic DNA typing: Methodology*, Academic Press, Waltham, MA, 2011; (*b*) H. A. Hammond, L. Jin, Y. Zhong, C. T. Caskey and R. Chakraborty, *Am. J. Hum. Genet.*, 1994, **55**, 175; (*c*) K. Sullivan, A. Mannucci, C. Kimpton and P. Gill, *BioTechniques*, 1993, **15**, 637.

12 D. Khodakov, L. Thredgold, C. E. Lenehan, G. G. Andersson, H. Kobus and A. V. Ellis, *Biomicrofluidics*, 2012, **6**, 26503.

- D.A. Khodakov, **A.S. Khodakova**, A. Linacre, A.V. Ellis.Toehold-mediated nonenzymatic DNA strand displacement as a platform for DNA genotyping. **Journal of The American Chemical Society, 2013**, 135(15), 5612-5619.

JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

Article

pubs.acs.org/JACS

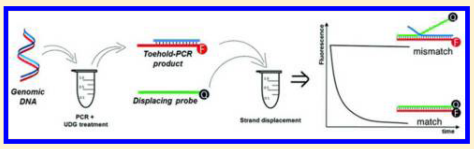
# Toehold-Mediated Nonenzymatic DNA Strand Displacement As a Platform for DNA Genotyping

Dmitriy A. Khodakov,*,† Anastasia S. Khodakova,‡ Adrian Linacre,‡ and Amanda V. Ellis*,†

†Flinders Centre for Nanoscale Science and Technology and ‡School of Biological Sciences, Flinders University, GPO Box 2100, Adelaide, S.A, 5001 Australia

Supporting Information

**ABSTRACT:** Toehold-mediated DNA strand displacement provides unique advantages in the construction and manipulation of multidimensional DNA nanostructures as well as nucleic acid sequence analysis. We demonstrate a step change in the use of toehold-mediated DNA strand displacement reactions, where a double-stranded DNA duplex, containing a single-stranded toehold domain, enzymatically generated and then treated as a molecular target for analysis. The approach was successfully implemented for human DNA genotyping, such as gender identification where the amelogenin gene was used as a model target system, and detecting single nucleotide polymorphisms of human mitochondrial DNA. Kinetics of the strand displacement was monitored by the quenched Förster resonance energy transfer effect.




## INTRODUCTION

DNA displacement reactions between double-stranded (ds)-DNA with strands of unequal length and single-stranded (ss)-oligonucleotides using a toehold structure, as a trigger point, enables DNA rehybridization in a fast "base-by-base" programmable controlled manner.[1,2] The advantage of this is that the route of strand displacement can easily be predicted due to the fact that the displacement is always known at the commencement of the displacement. Moreover, provided the entire sequence of the reacting oligonucleotides is known and taking into account all mismatches, such as point mutations, deletions, or insertions occurring during strand exchange, it is possible to predict the outcome and the kinetic and thermodynamic behavior of the displacement reaction.[3]

Since the discovery of toehold-mediated nonenzymatic DNA strand displacement reactions by Yurke et al.[4] in 2000 there has been a plethora of innovation in this new era of DNA nanotechnology.[5,6] Such innovations include the construction of artificial DNA nanoactuators,[7,8] computation DNA logic elements,[9−12] and 2-D and 3-D DNA nanoconstructions.[13,14] In particular, biosensing techniques based on strand displacement were targeted at synthetic nucleic acids,[3,15−20] proteins[21,22] and low molecular weight organic and inorganic molecules and ions.[15,23] However, to date, little attention has been paid to manipulating DNA strand displacement reactions for the analysis of real-life DNA samples, such as those of human, bacterial, or viral genomes.

Recently, a platform based on toehold-mediated strand displacement and atomic force microscopy (AFM) for label-free single nucleotide polymorphism (SNP) genotyping was described by Zhang et al.[16] The authors constructed a DNA-origami chip anchored with single-stranded capture probes which in turn were hybridized with partially complementary streptavidin labeled reporter probes. Application of fully complementary oligonucleotides (analytes) to the origami chip allowed the authors to perform toehold exchange of strands which were identified by the disappearance of streptavidin "white bulges" on the AFM images. This approach was subsequently applied to SNP typing using synthetic single-stranded target oligonucleotides, unrelated to real DNA diagnostics. A similar approach was demonstrated by Subramanian et al.[19] In another approach Picuri et al.[15] focused on the creation a universal translator for nucleic acid diagnostics. A toehold-mediated DNA strand displacement reaction was applied to the polymerase chain reaction (PCR) independent translation of biologically relevant DNA and RNA sequences into unique unrelated DNA sequences. This in turn could be recognized by another generic technique of choice. The technology was then used to identify synthetic oligonucleotides which mimic sequences of hepatitis C (HCV), influenza, and chicken pox viruses.

Previous work in literature[16−18,24] has predominately focused on short artificial ss-oligonucleotides as targets for identification, where the ds-DNA duplex containing a ss-toehold domain acted as a molecular probe. Moreover, the concentrations of the targets were well beyond those conventionally observed in real-life nucleic acid analyses. This in turn imposes strict limitations on the application of these techniques for the analysis of real DNA samples.

Seminal work by Pourmand et al.[25] described the analysis of human DNA length polymorphisms (short tandem repeats (STRs)). In this work the DNA strand displacement phenomenon was used as an auxiliary technique to remove

Received: November 7, 2012
Published: April 3, 2013





dx.doi.org/10.1021/ja310991r | J. Am. Chem. Soc. 2013, 135, 5612−5619

imperfectly hybridized (overhanging) ss-DNA from specific capture probes immobilized on a microarray surface. This approach consisted of three nested PCRs in order to obtain a ss-DNA target. Three consecutive hybridization procedures were then used to hybridize the ss-DNA target with an immobilized probe and remove all imperfectly hybridized ds-DNA, likely restricting this method for clinical or forensic applications.

Herein, we describe a new simpler approach for the analysis of real-life DNA samples using a toehold-mediated DNA strand displacement reaction where a partially complementary duplex, with a ss-toehold sequence, was used as a target for analysis. The approach consisted of two stages (Figure 1): (i) PCR amplification was used to generate a PCR product with a contiguous ss-toehold domain (further named toehold-PCR product). This was achieved using a substitution of deoxythymidine for deoxyuracil (dT→dU) in one of the PCR primers. Using this dU substituted primer then resulted in a PCR product with incorporated uracil bases. By treating the PCR product with uracil-DNA glycosylase (UDG) short ss-oligonucleotides were released from the 5′ of the PCR product, making them suitable for toehold-mediated strand displacement reactions; and (ii) toehold-mediated DNA strand displacement reactions between the target toehold-PCR product and single-stranded chemically synthesized oligonucleotide molecular probes were performed using a quenched Föster resonance energy transfer (FRET) technique to monitor how the reaction proceeded. The analytical sensitivity of the entire approach is determined by the sensitivity of the first PCR stage and could potentially reach a level as little as 5−10 DNA molecules.[26] The amelogenin gene, commonly used both in forensic and medical applications for human gender identification, was used as a model system for human DNA genotyping.[27,28] Using this model system the enzymatic generation of the toehold-PCR products of the male and female human genome, with subsequent strand displacement assisted discrimination, was shown. Moreover, the developed approach was directly adapted for SNP testing. As an example, a C-to-T (C16223T) single nucleotide substitution at the 16223 position of the hyper-variable region 1 (HVR-1) of the mtDNA found in one of the authors DNA samples was discriminated. It is envisaged that this model system could be easily adapted to other DNA genotyping molecular diagnostics.

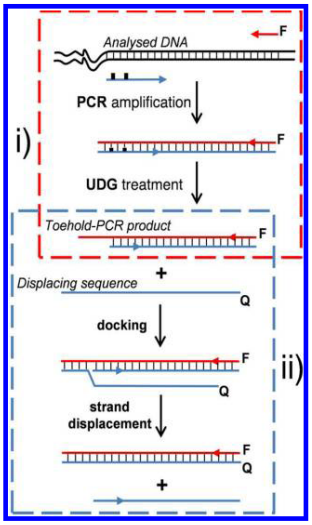



**Figure 1.** Schematic representation of the approach proposed for the analysis of real-life ds-DNA samples using toehold-mediated strand displacement reaction. (i) Deoxyuracil (■) modified forward and fluorophore (F) labeled reverse primers are used for the amplification of the target DNA. Following treatment of the dU modified amplification product with uracil-DNA glycosylase a ds-PCR product containing a ss-toehold sequence (toehold-PCR product) was generated. (ii) Toehold-mediated strand displacement between the toehold-PCR product and a chemically synthesized displacing sequence labeled with quencher (Q) oligonucleotide molecular probes. Displacement is monitored using a quenched FRET technique.

## MATERIALS AND METHODS

**Nucleic Acid Isolation and PCR Amplification.** Human genomic and mitochondrial DNA were coextracted from the authors' own blood samples using a QIAamp DNA blood mini kit (Qiagen, Germany), according to the recommendations of the manufacturer. PCR primers, including dU modified and fluorescently labeled primers (FAM and HEX) (see Table 1) were purchased from IDT DNA Technology, USA, and used at a concentration of 0.2 $\mu$M. Melting temperatures were calculated using Oligo Analyzer 3.1 (IDT DNA Technology, USA) under the following conditions: an oligonucleotide concentration of 0.2 $\mu$M, $Na^+$ concentration of 100 mM, $Mg^{2+}$ concentration of 2.5 mM, and deoxyribonucleotide triphosphates (dNTPs) concentration of 0.8 mM. Both amelogenin and SNP targeted PCRs were performed using a HotStar Taq DNA polymerase (Qiagen, Germany) within 1× HotStar Taq PCR buffer with a final $MgCl_2$ concentration of 2.5 mM. The concentration of the standard deoxynucleotide triphosphates (dATP, dTTP, dGTP, dCTP) was 0.2 mM. A PCR amplification regime of 95 °C for 15 min, 30 cycles of 94 °C for 20 s, 61 °C for 30 s, 72 °C for 30 s, and a final elongation of 72 °C for 3 min was used. A template of 5 ng of the extracted DNA samples was used.

Directly after the PCR, uracil-DNA glycosylase (NEB, USA) (2.5 U) was added to the PCR solution, and the entire mixture was then gently mixed via pipetting. After incubation at room temperature for 5 min the mixture was heated to 95 °C for 5 min and then cooled back to room temperature. The PCR-UDG mixtures were purified using a Qiaquick PCR purification kit (Qiagen, Germany) for the amelogenin system and Amicon Ultra-0.5 30K (Millipore, USA) for SNP testing, according to the recommendations of the manufacturers. The nucleic acids were then eluted from the Qiaquick PCR purification kit, or recovered Amicon Ultra-0.5 30K, with a displacement TEM buffer (pH 8), consisting of Tris·HCl (10 mM), EDTA (1 mM), and $MgCl_2$ (12.5 mM). Quantification of the eluted PCR products was carried out using a NanoDrop 1000 spectrophotometer (Thermo Scientific, USA). Capillary electrophoresis (CE) analysis was carried out on a ABI-3130XL Genetic Analyzer (Life Technologies, USA), using a GeneScan 500 LIZ size standard (Life Technologies, USA).

**Strand Displacement Reaction.** Displacing sequences (X*i*, Y*i*, SNP-C*i*, and SNP-T*i*, see Table 1) labeled at the 3′ termini with the fluorescent quencher carboxytetramethylrhodamine (TAMRA), for FRET analysis, were purchased from Eurogentec, Belgium.

All displacement reactions were carried out at 30 °C, if not mentioned otherwise, in a reaction volume (15 $\mu$L) using a RotorGene 3000 real-time PCR thermocycler. Prior to the displacement reaction the toehold-PCR product (2 pMole, dissolved in 1× TEM buffer) was placed in a thin-wall PCR tube (0.2 mL) and briefly centrifuged. Then, the displacing sequence (20 pMole, dissolved in 2 $\mu$L of 1× TEM buffer), if not mentioned otherwise, was carefully applied into the lid of the same PCR tube. The tube lid was then carefully closed and the tube placed into the thermocycler carefully avoiding mixing of the


dx.doi.org/10.1021/ja310991r | J. Am. Chem. Soc. 2013, 135, 5612−5619

**Table 1. Sequences and Melting Temperatures of Primers and Displacing Oligonucleotides**

| | Sequence | $T_m$, °C |
|---|---|---|
| | PCR Primers | |
| AMEL-F-dU-4 | 5′- CCC *dU*GG GCT CTG TAA AGA ATA GTG-3′ | 65.3 |
| AMEL-F-dU-9 | 5′- CCC TGG GC*dU* CTG TAA AGA ATA GTG-3′ | 65.3 |
| AMEL-F-dU-4,9 | 5′- CCC *dU*GG GC*dU* CTG TAA AGA ATA GTG-3′ | 65.3 |
| AMEL-F-HEX -dU-4,9 | 5′-HEX-CCC *dU*GG GC*dU* CTG TAA AGA ATA GTG-3′ | 65.3 |
| AMEL-FAM-R | 5′-FAM-ATC AGA GCT TAA ACT GGG AAG CT-3′ | 65.1 |
| HVR-F-dU-5,9 | 5′-CTA G*dU*G GG*dU* GAG GGG TGG CT-3′ | 68.0 |
| HVR-FAM-R | 5′-FAM-ATGCTTACAAGCAAGTACAGCAAT-3′ | 64.1 |
| | Displacing Sequences | |
| X*i* (106) | 5′-CCCTGGGCTCTGTAAAGAATAGTGTGTTGATTCTTTATCCCAGAT - - - - - - GTTT CTCAAGTGGTCCTGATTTTACAGTTCCTACCACCAGCTTCCCAGTTTAAGCTCTGAT-TAMRA-3′ | |
| Y*i* (112) | 5′-CCCTGGGCTCTGTAAAGAATAGTGGGTGGATTCTTCATCCCAAATAAAGTGGTTT CTCAAGTGGTCCCAATTTTACAGTTCCTACCATCAGCTTCCCAGTTTAAGCTCTGAT-TAMRA-3′ | |
| SNP-C*i* | 5′-CTAGTGGGTGAGGGGTGGCTTTGGAGTTGCAGTTGATGTGTGATAGTTGAGGGTT GATTGCTGTACTTGCTTGTAAGCAT-TAMRA-3′ | |
| SNP-T*i* | 5′-CTAGTGGGTGAGGGGTGGCTTTGGAGTTGCAGTTGATGTGTGATAGTTGAAGGTT GATTGCTGTACTTGCTTGTAAGCAT-TAMRA-3′ | |

toehold-PCR product and the displacing sequence before the run protocol started (for details see Supporting Information, S2). As controls the UDG untreated PCR products and the single FAM labeled reverse primer (R-FAM) were used.

Acquisition of fluorescent signals was performed within the SYBR Green/FAM channel at a gain of 10 within a time interval of 20 s between fluorescent reads. Normalization of the raw fluorescent (see Figure S5) signals was made by dividing the signal values by the initial signal value, as described previously.[2]

Curve-fitting of the kinetic data was carried out using the nonlinear curve fitting function in OriginPro 8 software (Origin Corporation, USA). The data were fit to the first-order equation as described by Baker et al.[29]

## ■ RESULTS AND DISCUSSION

**Model System/Target.** The amelogenin gene was chosen as a target model system because the Y chromosome has a six base pair insertion compared to the homologue on the X chromosome. Typically, PCR amplification of intron 1 of the amelogenin locus with specific primers produces a homogeneous PCR product for female samples (two X chromosomes, XX), while producing a heterogeneous product for male DNA samples (X and Y chromosomes, XY).[30]

**Primer Design.** A DNA duplex with a ss-toehold domain sequence was required for the toehold-mediated DNA strand exchange. In order to achieve this, the forward PCR primer had the deoxythymidine substituted for deoxyuracil (dT→dU). This resulted in the generation of a PCR product with incorporated dU bases at the primer location. Consecutive treatment of the PCR mixture with uracil-DNA glycosylase, an enzyme which removes uracil bases from DNA by glycosidic bond hydrolysis[31] and heating of the reaction to 95 °C lead to the hydrolysis of the DNA phosphate backbone producing the PCR products with a ss-toehold domain (Figure 1i).

A primer set with sequences based on those from Power Plex 16 System (Promega, USA),[32] for amplification of the amelogenin gene, was used. These are primers that are typically used in the majority of forensic PCR systems for human STR analysis, except with the dT→dU substitutions in the sequence of the forward primer. Three different dU substituted forward primers (Table 1) were used in order to investigate the influence of toehold sequence length on the efficiency of the displacement. The primers F-dU-4 and F-dU-9 were generated with one dT→dU substitution at the fourth and ninth position (counting from the 5′ termini), respectively. These substitutions therefore resulted in the formation of PCR products with 4 and 9 nucleotide toehold lengths. In the primer AMEL-F-dU-4,9, dT→dU substitutions were made at two positions to provide a 9 nucleotide toehold length as well. This was achieved by removing the uracil bases from both the fourth and ninth positions simultaneously.

The reverse primer (AMEL-R-FAM) for PCR amplification was labeled with FAM fluorescent dye at its 5′ termini, which was then used as a reporter fluorescent dye for monitoring reaction kinetics using the quenched FRET effect (Figure 1ii).

In order to evaluate the efficiency of the PCR amplification with dU modified primers and subsequent UDG assisted generation of a ss-toehold sequence, the AMEL-F-dU-4,9 primer was labeled with HEX fluorescent dye (AMEL-F-HEX-dU-4,9) at its 5′ termini in order to be observed using CE.

**Amelogenin Gene PCR Amplification and Formation of a Toehold Sequence.** For the PCR amplification with dU modified primers special attention was required in the selection of the DNA polymerase due to the fact that native DNA polymerases isolated from archaea, for example, Pfu DNA polymerase (Promega, USA), Pfx DNA polymerase (Life Technologies, USA) or Phusion DNA polymerase (NEB, USA), bind strongly to the uracil-containing DNA template and stall further polymerization.[33] For this reason only DNA polymerases originating from bacterial host strains or a mutant Pfu DNA polymerase, for example, as described by Norholm,[34] should be used while working with dU modified primers. Therefore, HotStar Taq DNA polymerase (Qiagen, Germany), isolated from *Thermus aquaticus*, was used in this research.

The analysis of the PCR amplification using dU substituted primers and the subsequent UDG treatment was then performed. To achieve this the PCR amplification of human both female and male DNA samples (5 ng, each) using AMEL-F-HEX-dU-4,9 and AMEL-R-FAM primers was conducted. Immediately after the PCR, the mixtures were treated with UDG. The results of the reactions were then confirmed using CE (see Supporting Information, S1). Figures S1A, S1B, S3A, and S3B show the results of the CE analysis of the PCR products before the UDG treatment. The peaks at the positions (approximately) of the 106 and 112 nucleotides belong to X and Y chromosomes, respectively. It can be seen that the peaks corresponding to the HEX fluorescent dye (Figures S2B and


dx.doi.org/10.1021/ja310991r | J. Am. Chem. Soc. 2013, 135, 5612−5619

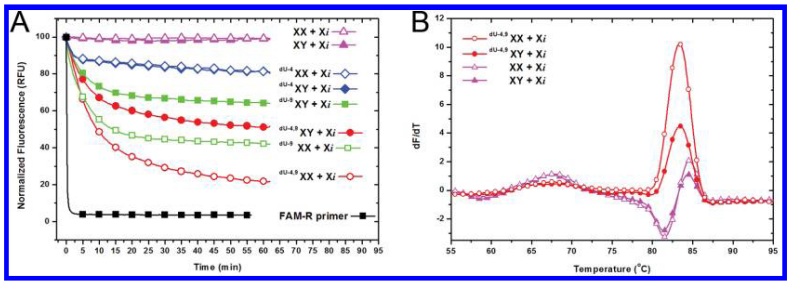



**Figure 2.** (A) Effect of the toehold length on the efficiency of the strand displacement reaction between the amelogenin PCR products (XX, female = hollow symbols; XY, male = solid symbols) and X*i* displacing sequence. Toehold lengths of 4 ($^{dU-4}$toehold-PCR products, ◆ and ◇) and 9 nucleotides ($^{dU-9}$- and $^{dU-4,9}$toehold-PCR products, □ and ■ and ○ and ●, respectively) were investigated. The female (XX, △) and male (XY, ▲) PCR products with no toehold sequence as well as the fluorescein labeled reverse primer (FAM-R, solid squares) were used as controls. (B) Melting temperature analysis of the corresponding displaced products.

S4B) disappear completely after the UDG treatment, indicating that the phosphate backbone at the site of the uracil base had been destroyed.

Six different fluorescently labeled amelogenin toehold-PCR products were obtained via amplification of male and female DNA with three different primer pairs and treatment with UDG. Every primer pair consisted of the AMEL-R-FAM primer and each of three dU substituted primers (AMEL-F-dU-4, AMEL-F-dU-9, and AMEL-F-dU-4,9) (Table 1).

**DNA Strand Displacement Monitoring Using FRET.** Two chemically synthesized displacing sequences labeled at their 3′ termini with TAMRA (for the FRET quenching effect) (X*i* and Y*i*, shown in Table 1) were designed for reacting with the FAM fluorescently labeled amelogenin toehold-PCR products. The sequence of the X*i* displacing sequence, having a length of 106 nucleotides, was fully complementary to the longer strand of the toehold-PCR product produced from the X chromosome, while the Y*i* displacing sequence, having a length of 112 nucleotides, was fully complementary to the longer strand of the toehold-PCR product of the Y chromosome.

The difference in length between the X*i* and Y*i* displacing sequences was attributed to a 6 nucleotide insertion (AAAGTG) at the 46th position (counting from the 5′ nucleotide of the forward primer). Moreover, addition of single nucleotide substitutions namely T25G, T28G, T36C, G43A, T68C, G69A, and C88T, corresponding to naturally occurring polymorphisms in the X and Y chromosomes of the human genome,[30] was also present in the X*i* and Y*i* displacing sequences. According to literature[2,3,15,19] the influence of these polymorphisms is controversial, but in this particular case, this may have negligible influence on the differentiation between X and Y chromosomes compared to the insertion of six nucleotides.

A 10-fold excess of the TAMRA labeled X*i* sequence was added to both the fluorescently labeled female and male PCR products and the FRET quenching effect monitored. The results of the experiments are shown in Figure 2A. The female (XX) and male (XY) UDG untreated PCR product control samples (Figure 2A; Δ, ▲) showed the degree of conversion to be approximately 2% (±3%), indicating that no reaction occurred.

Among the three toehold-PCR products obtained after the UDG treatment the highest degree of conversion of the strand displacement was observed for the amelogenin toehold-PCR product, generated using the AMEL-F-dU-4,9 forward primer ($^{dU4,9}$toehold-PCR product) with two dT→dU substitutions. The displacement levels in this case reached 79% for the female toehold-PCR product ($^{dU-4,9}$XX, Figure 2A, ○) and 49% for the male toehold-PCR product ($^{dU-4,9}$XY, Figure 2A, ●).

Comparison was then performed between the degree of conversion for the amelogenin toehold-PCR products made with AMEL-F-dU-9 and AMEl-F-dU-4,9 primers ($^{dU-9}$XX and $^{dU-4,9}$XX, respectively) reacting with X*i*, (Figure 2A). Although the toehold domains in both cases theoretically have the same length (9 nucleotides), the degree of displacement for the reaction with the toehold-PCR product produced with the F-dU-9 primer was noticeably lower and found to be around 58% (±4) and 37% (±3) for female $^{dU-9}$XX (Figure 2A, □) and male $^{dU-9}$XY samples (Figure 2A, ■), respectively. This could best be explained by incomplete dissociation of the digested 5′ sequence consisting of 9 nucleotides and having a melting temperature of 45.3 °C. Most likely this nondissociated sequence acts as a protector for toehold strand displacement.[35] This indicated that a single dU substitution was insufficient to provide a fully unprotected toehold domain of 9 nucleotides in length. Thus, the AMEL-F-dU-9 primer was excluded from further investigation.

The amelogenin $^{dU4}$toehold-PCR product with the shortest toehold length of 4 nucleotides gave the lowest degree of displacement at around 18 ± 3% for both female ($^{dU-4}$XX Figure 2A, ◇) and male ($^{dU-4}$XY Figure 2A, ◆) samples. That in turn indicates that, in this particular case, the toehold consisting of 4 nucleotides ($T_m$ < 10 °C) was not able to initiate an effective strand displacement process.[1] As a result the AMEL-F-dU-4,9 primer proved to be the most efficient for the generation of a ss-toehold domain in a PCR product using the proposed method.

Since the displacement reactions were carried out using a real-time PCR thermocycler, this provided the opportunity to perform melting temperature analysis of the displaced products. The melting analysis (Figure 2B) shows the major melting peak of the displaced product at 83.2 °C (79.7 °C calculated) for female (Figure 2B, ○) and male (Figure 2B, ●) PCR products produced with the AMEL-F-dU-4,9 primer. The area under the curve for these samples differs by approximately a factor of 1.87, which closely corresponds to the ratio between of the amount of the X chromosome in female DNA and in male DNA.


dx.doi.org/10.1021/ja310991r | J. Am. Chem. Soc. 2013, 135, 5612−5619

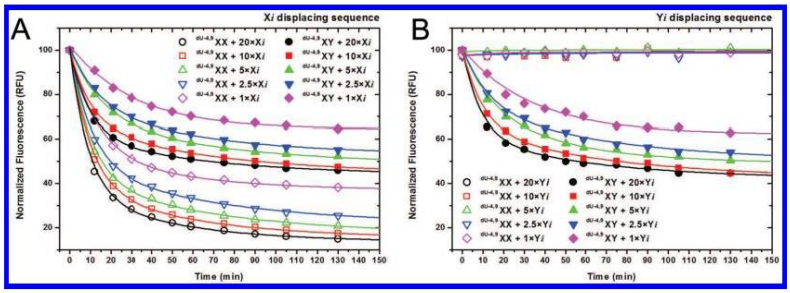



**Figure 3.** Effect of the displacing sequence excess on the degree of strand displacement conversion. Strand displacement reaction between the amelogenin $^{dU-4,9}$toehold-PCR products containing a 9 nucleotides long ss-toehold sequence (female $^{dU-4,9}$XX = hollow symbols, male $^{dU-4,9}$XY = solid symbols) and 20× (circles), 10× (squares), 5× (up-pointing triangles), 2.5× (down-pointing triangles), and 1× (diamonds) excesses of the X*i* (A) and Y*i* (B) displacing sequences.

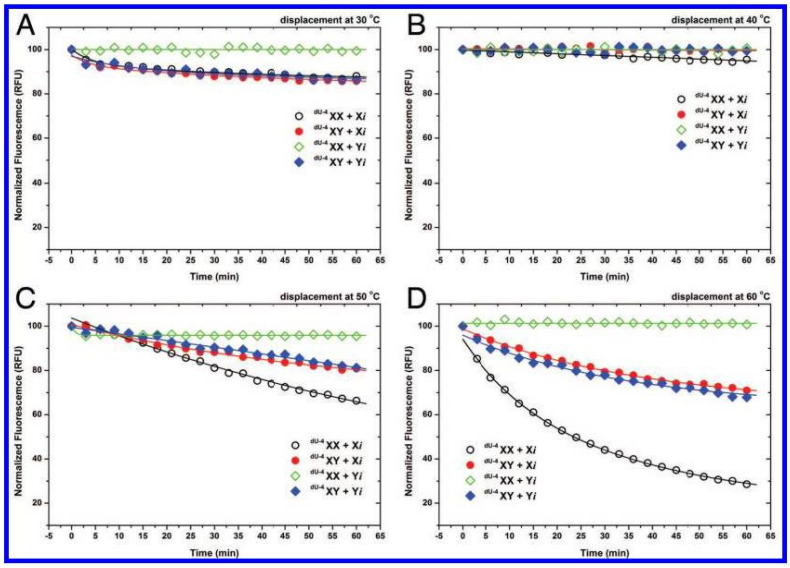



**Figure 4.** Kinetics of the displacement reactions between the amelogenin $^{dU-4}$toehold-PCR products and a 10-fold excess of X*i* (circles) and Y*i* (diamonds) displacing sequences performed at the different temperatures: (A) 30 °C, (B) 40 °C, (C) 50 °C, and (D) 60 °C.

The melting curves of the UDG untreated amelogenin samples (no toehold structure was created) showed a different melting behavior (Figure 2B, Δ and ▲). Such behavior represents a dissociative pathway of the DNA strand displacement which occurs at a temperature close to melting.[36] Similar melting trends were observed for the $^{dU4}$toehold-PCR products with 4 base length toeholds (data not shown).

To further optimize the efficiency of the strand displacement, the ratio (excess) of the displacing sequence to the amelogenin toehold-PCR products was evaluated. Five different ratios of both X*i* and Y*i* sequences, in the range of a 20-fold excess (20×) to a 1-fold excess (1×), were used. Figure 3A shows the difference between the degree of conversion for both female ($^{dU4,9}$XX, Figure 3A, hollow symbols) and male ($^{dU4,9}$XY, Figure 3A, solid symbols) samples reacting with the female displacing sequences (X*i*). In all cases there was a gradual decrease in the degree of conversion. For female samples this degree of conversion was 4.3 ± 1.1% and for male samples this was 3.7 ± 1.1% across the 20×, 10×, 5×, and 2.5× X*i* displacement sequence excess concentrations. However, the samples with equimolar ratio (1×) of the X*i* displacing sequence to the toehold-PCR products (Figure 3A, ◇ female and ◆ male toehold-PCR products) differ from the 2.5× excess by 9.2 ± 1.2% for female samples and 8.3 ± 1.4% for male samples. This difference can be attributed to either incorrect determination of nucleic acid concentration or the side reactions of the displacing sequence with traces of unreacted AMEL-R-FAM primer. However, these processes are rendered negligible when using larger excesses of the displacing sequences.


dx.doi.org/10.1021/ja310991r | J. Am. Chem. Soc. 2013, 135, 5612−5619

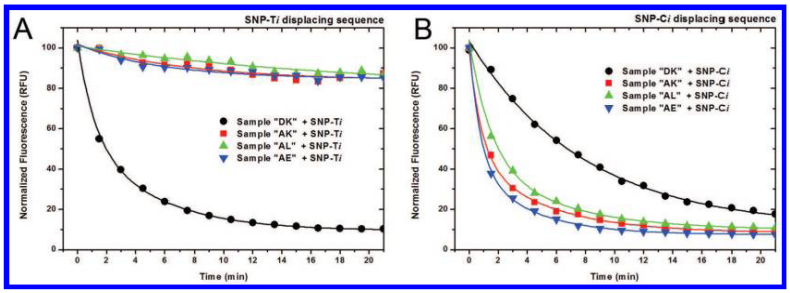



**Figure 5.** mtDNA C16223T SNP discrimination (A) with the displacing sequence carrying the dA nucleotide at the opposite position to the SNP, (B) with the displacing sequence carrying the dG nucleotide an the opposite position to the SNP. Sample "DK" (circles) has a dT nucleotide at the 16223 position, while samples "AK", "AL", and "AE" ( ■, ▲, ▼, respectively) have a dC nucleotide at the 16223 position.

Figure 3B shows the male displacing sequence (*Yi*) reacting with both amelogenin female ($^{dU4,9}$XX, Figure 3B, hollow symbols) and male ($^{dU4,9}$XY, Figure 3B, solid symbols) toehold-PCR products. Since the female toehold-PCR product (XX) did not contain a sequence that was fully complementary to the *Yi* displacing sequence then the reaction between the *Yi* and the female toehold-PCR product did not occur at any excess of the displacing sequence. On the other hand the kinetics of the strand displacement between the male toehold-PCR product ($^{dU4,9}$XY) and the *Yi* displacing sequence (Figure 3B, solid symbols) was similar to that observed for the reaction with the *Xi* displacing sequence (Figure 3A, solid symbols). All of the evaluated ratios (excesses) were successful at discriminating both female and male toehold-PCR samples within the first 20 min of the displacement reaction. The 10-fold excess (10×) was then chosen for all consecutive experiments.

It was found that by conducting the strand displacement at elevated temperatures, the reaction was noticeably accelerated. The reaction kinetics for the amelogenin $^{dU-4,9}$XX and $^{dU-4,9}$XY toehold-PCR products were studied at four temperatures of namely, 30, 40, 50, and 60 °C (Figure S6). The kinetic data obtained were fitted to the first-order reaction equation according to Baker et al.[29] and Reynaldo et al.[36] using OriginPro 8 software (Origin Corporation, USA). The observed rate constants ($k_{obs}$) for the displacement reactions are summarized Table S1. Despite the fact that the research presented here uses reasonably long DNA duplexes and displacing sequences, the observed rate constant values ($\sim 10^{-3}$ $s^{-1}$) were orders of magnitude faster than the values reported for short oligonucleotides immobilized on a surface ($\sim 10^{-4}$ $s^{-1}$, Baker et al.)[29] or dissolved in solution ($\sim 10^{-6}$ $s^{-1}$, no toehold-mediated displacement was investigated, Reynaldo et al.).[36] The reaction kinetics for the amelogenin $^{dU-4}$XX and $^{dU-4}$XY toehold-PCR products were also then studied at the same four temperatures as those used for the amelogenin $^{dU-4,9}$toehold-PCR products (Figure 4). Interesting, at 30 °C Figure 4A shows an undistinguishable degree of displacement at 13% for the amelogenin female toehold-PCR product ($^{dU-4}$XX), reacting with *Xi*, and the male toehold-PCR product ($^{dU-4}$XY), reacting with both *Xi* and *Yi* (Figure 4A ○, ●, ◆, respectively). As expected, the reaction of the female toehold-PCR product $^{dU-4}$XX with *Yi* displacing sequence showed no conversion at all (Figure 4A ◇). Similarly, at 40 °C there was no observed displacement for the male toehold-PCR product ($^{dU-4}$XY) reacting with both the *Xi* and *Yi* displacing sequences (Figure 4B ●, ◆, respectively). However, for the female toehold-PCR product ($^{dU-4}$XX) a lower degree of conversion of 3 ± 2% was observed with the *Xi* displacing sequence (Figure 4B ○). This fall in conversion (from 13% to 3%) may be explained by the reduced stability of the 4 nucleotide toehold structure at 40 °C. Increasing the reaction temperature to 50 °C and then to 60 °C (Figure 4C,D, respectively) resulted in distinguishable displacements for both amelogenin female and male toehold-PCR products allowing for their discrimination. Since the formation of a toehold structure at these temperatures is unlikely, it is most probable that the strand displacement reactions are activated by means of partial melting of the toehold-PCR products.[36] Moreover, the observed rate constants for the reaction kinetics, shown in Figure 4C,D, were lower compared to those observed for the $^{dU-4,9}$toehold-PCR products (Table S1 and Figure S6).

Activation energies ($E_a$) (Table S1) were calculated from the slopes of the Arrhenius plots for the reactions of $^{dU-4,9}$XX with *Xi*, $^{dU-4,9}$XY with *Xi*, $^{dU-4,9}$XY with *Yi*, and $^{dU-4}$XX with *Xi* toehold-PCR products (Figure S7 ○, ●, ◆). Table S1 shows that $E_a$'s of the reactions of the amelogenin PCR products containing the 9 nucleotide long toehold domain (Figure S7 ○, ●, ◆) have a similar value (17.1 ± 0.95 kcal/mol). However, the $E_a$ value (36.9 kcal/mol) for the displacement reaction of the $^{dU-4}$XX toehold-PCR product with *Xi* was approximately twice as high as the displacement reaction of the $^{dU-4,9}$XX toehold-PCR product with *Xi* ($E_a$ = 17.2 kcal/mol). This suggests that in the case of the reaction between the $^{dU-4}$XX toehold-PCR product with *Xi*, the dissociative activation of the displacement occurs at elevated temperatures. On the other hand, the displacement reaction between both the female $^{dU-4,9}$XX and the $^{dU-4}$XX toehold-PCR products and *Yi* displacing sequence (Figure S6 ◇, and Figure 4 ◇, respectively) did not occur at any of the temperatures tested.

**SNP Testing.** SNPs are the predominant variant in the human genome, and their detection plays a pivotal role in medical diagnostics, prediction of treatment, and the outcome of genetically determined diseases. Typically, SNPs are distributed throughout the human genome, including the mitochondrial genome. Two hypervariable regions (HVR-I and HVR-II) within the mitochondrial DNA (mtDNA) contain an abundance of SNP markers which provide highly useful information for determining human maternal ancestry.[37,38] Four DNA samples of the authors' own blood specimens were sequenced within the positions 16106−16339 of the HRV-1


dx.doi.org/10.1021/ja310991r | J. Am. Chem. Soc. 2013, 135, 5612−5619

region (according to the Gene Bank Acc. no. J01415.2; details in Supporting Information, S6). The DNA sample "DK" contained substitution C-to-T at the position 16223. In order to distinguish the "DK" sample from others, a new set of HRV-1 specific dU and FAM modified primers and two displacing sequences (labeled with TAMRA at 3′ termini) was designed (Table 1). The dU modified primer contained two dU nucleotides at the fifth and ninth position providing a toehold length of 9 nucleotides within the total length of the PCR product (80 base pairs). The displacing sequences, SNP-C*i* and SNP-T*i*, contained dG and dA nucleotides at the 51 position, respectively, which are able to distinguish the C16223T substitution in the analyzed DNA samples. After PCR amplification the dU modified PCR products were treated as per the amelogenin system described previously. The displacement reaction was performed at 30 °C with 10-fold displacement sequences excess. Figure 5 shows the results of the displacement. The "DK" toehold-PCR product, with the C16223T substitution, was easily discriminated from the other three authors' samples containing no substitutions at the position 16223 using the SNP-T*i* displacing sequence (Figure 5A). The "DK" sample showed a displacement level of ~90.0 ± 5.3% after 20 min (Figure 5A ●), while the others – only 13.5 ± 6.7% (Figure 5A ■, ▲, ▼). Figure 5B shows the displacing reaction using the SNP-C*i* displacing sequence. It can be seen that within the first 10 min of the reaction, the "DK" sample can be reliably distinguished from the other three samples. However, the longer incubation time leads to a almost common plateau for all four analyzed samples. These results are in full accordance with those showed by Picuri et al.,[15] where the authors correlated the difference in discrimination ability of the C-to-T substituted displacing sequence with a secondary structure located throughout the toehold domain.[15] However in our case no hindered toehold domain structures, calculated using Mfold web server,[39] were found for both displacing sequences and this phenomenon requires further investigation.

## ■ CONCLUSION

DNA strand exchange processes have recently been introduced as a major technique for DNA nanostructuring, machinery, computation, and biosensing. A "zipper" mode or a consecutive "base-by-base" rehybridization mechanism allows for the highly efficient discrimination of reacting ds-DNA. However, DNA diagnostic related work in this area, thus far, has almost exclusively focused on the analysis of comparatively high concentrations of synthetic single-stranded nucleic acids. Here, we successfully demonstrate a new approach to the use of the strand displacement reaction for the analysis of ds-DNA using the amelogenin gene as a model system for gender discrimination. PCR products with length of 106 and 112 bp of the amelogenin locus on the X and Y chromosomes, respectively, of this gene were prepared with a ss-toehold sequence which were then subjected to strand displacement using synthetic complementary oligonucleotides. It was found that the ss-toehold domain consisted of 9 nucleotides allowed the discrimination between male and female samples in <10 min of the strand displacement reaction conducted at 30 °C. By increasing the reaction temperature to 60 °C, the displacement level >70% was achieved within 1 min. The use of a toehold domain with a length of 4 nucleotides was only found to be successful at discrimination when using elevated temperatures of 50 and 60 °C.

In addition the approach was demonstrated for the SNP genotyping of real-life ds-DNA samples using mtDNA as a target. The C-to-T substitution in 80 bp PCR products was discriminated with a displacing sequence carrying a dA nucleotide at the opposite position to the substitution. However, the displacing sequence with a dG nucleotide allowed only kinetic discrimination within first several minutes of the reaction. Again, our data demonstrate that the current knowledge of SNP influence on the efficiency of toehold-mediated strand displacement is still controversial and requires additional investigations.

In summary, a novel genotyping approach has been developed which requires only a simple substitution of deoxythymidine for deoxyuracil in one of the existing PCR primers and UDG treatment followed by PCR. This makes this method easily applicable to most DNA genotyping systems. The approach is especially useful for systems which deal with the analysis of nucleic acids with similar hybridization efficiencies and those prone to cross-hybridization.[22] Finally, the strategies described herein are directly adaptable to microarray technologies.

## ■ ASSOCIATED CONTENT

### Supporting Information

CE assessment of the efficiency of the PCR amplification using dU substituted primers with UDG treatment. Strand displacement real-time monitoring workflow. An example of raw fluorescent data. Kinetic data of the strand displacement reactions. DNA sequencing. This material is available free of charge via the Internet at http://pubs.acs.org.

## ■ AUTHOR INFORMATION

### Corresponding Author

dmitriy.khodakov@flinders.edu.au; amanda.ellis@flinders.edua.au

### Notes

The authors declare no competing financial interest.

## ■ REFERENCES

(1) Zhang, D. Y.; Winfree, E. *J. Am. Chem. Soc.* **2009**, *131*, 17303–17314.
(2) Genot, A.; Zhang, D.; Bath, J.; Turberfield, A. *J. Am. Chem. Soc.* **2011**, *133*, 2177–2182.
(3) Zhang, D. Y.; Chen, S. X.; Yin, P. *Nat. Chem.* **2012**, *4*, 208–14.
(4) Yurke, B.; Turberfield, A. J.; Mills, A. P.; Simmel, F. C.; Neumann, J. L. *Nature* **2000**, *406*, 605–608.
(5) Teller, C.; Willner, I. *Curr. Opin. Biotechnol.* **2010**, *21*, 376–391.
(6) Zhang, D. Y.; Seelig, G. *Nat. Chem.* **2011**, *3*, 103–113.
(7) Yan, H.; Zhang, X.; Shen, Z.; Seeman, N. C. *Nature* **2002**, *415*, 62–65.
(8) Turberfield, A.; Mitchell, J.; Yurke, B.; Mills, A.; Blakey, M.; Simmel, F. *Phys. Rev. Lett.* **2003**, *90*, 1–4.
(9) Frezza, B.; Cockroft, S.; Ghadiri, M. *J. Am. Chem. Soc.* **2007**, *129*, 14875–14879.
(10) Shin, J.; Pierce, N. *Nano Lett.* **2004**, *4*, 905–909.
(11) Seelig, G.; Soloveichik, D.; Zhang, D. Y.; Winfree, E. *Science* **2006**, *314*, 1585–1588.
(12) Shlyahovsky, B.; Li, Y.; Lioubashevski, O.; Elbaz, J.; Willner, I. *ACS Nano* **2009**, *3*, 1831–1843.
(13) Andersen, E. S.; Dong, M.; Nielsen, M. M.; Jahn, K.; Subramani, R.; Mamdouh, W.; Golas, M. M.; Sander, B.; Stark, H.; Oliveira, C. L. P.; Pedersen, J. S.; Birkedal, V.; Besenbacher, F.; Gothelf, K. V; Kjems, J. *Nature* **2009**, *459*, 73–76.
(14) Tian, Y.; Mao, C. *J. Am. Chem. Soc.* **2004**, *126*, 11410–11411.

(15) Picuri, J. M.; Frezza, B. M.; Ghadiri, M. R. *J. Am. Chem. Soc.* **2009**, *131*, 9368−9377.
(16) Zhang, Z.; Zeng, D.; Ma, H.; Feng, G.; Hu, J.; He, L.; Li, C.; Fan, C. *Small* **2010**, *6*, 1854−1858.
(17) Tian, Y.; Mao, C. *Talanta* **2005**, *67*, 532−537.
(18) Wang, D.; Tang, W.; Wu, X.; Wang, X.; Chen, G.; Chen, Q.; Li, N.; Liu, F. *Anal. Chem.* **2012**, *84*, 708−714.
(19) Subramanian, H.; Chakraborty, B.; Sha, R.; Seeman, N. C. *Nano Lett.* **2011**, *11*, 910−913.
(20) Li, B.; Ellington, A. D.; Chen, X. *Nucleic Acids Res.* **2011**, *39*, e110.
(21) Duose, D. Y.; Schweller, R. M.; Hittelman, W. N.; Diehl, M. R. *Bioconjugate Chem.* **2010**, *21*, 2327−2331.
(22) Song, G.; Chen, M.; Chen, C.; Wang, C.; Hu, D.; Ren, J.; Qu, X. *Biochimie* **2010**, *92*, 121−127.
(23) Shen, Q.; Tang, S.; Li, W.; Nie, Z.; Liu, Z.; Huang, Y.; Yao, S. *Chem. Commun.* **2011**, *48*, 281−283.
(24) Li, Q.; Luan, G.; Guo, Q.; Liang, J. *Nucleic Acids Res.* **2002**, *30*, E5.
(25) Pourmand, N.; Caramuta, S.; Villablanca, A.; Mori, S.; Karhanek, M.; Wang, S. X.; Davis, R. W. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 6146−6151.
(26) Spangler, R.; Goddard, N. L.; Thaler, D. S. *PloS one* **2009**, *4*, e7010.
(27) Thangaraj, K.; Reddy, A. G.; Singh, L. *Int. J. Leg. Med.* **2002**, *116*, 121−123.
(28) Mannucci, A.; Sullivan, K. M.; Ivanov, P. L.; Gill, P. *Int. J. Leg. Med.* **1994**, *106*, 190−193.
(29) Baker, B. A.; Milam, V. T. *Nucleic Acids Res.* **2011**, *39*, e99.
(30) Sullivan, K. M.; Mannucci, A.; Kimpton, C. P.; Gill, P. *BioTechniques* **1993**, *15*, 636−641.
(31) Longo, M. C.; Berninger, M. S.; Hartley, J. L. *Gene* **1990**, *93*, 125−128.
(32) Krenke, B. E.; Tereba, A.; Anderson, S. J.; Buel, E.; Culhane, S.; Finis, C. J.; Tomsey, C. S.; Zachetti, J. M.; Masibay, A.; Rabbach, D. R.; Amiott, E. A.; Sprecher, C. J. *J. Forensic Sci.* **2002**, *47*, 773−785.
(33) Greagg, M.; Fogg, M. J.; Panayotou, G.; Evans, S. J.; Connolly, B. A.; Pearl, L. H. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 9045−9050.
(34) Nørholm, M. H. H. *BMC Biotechnol.* **2010**, *10*, 21.
(35) Zhang, D. Y. *J. Am. Chem. Soc.* **2011**, *133*, 1077−1086.
(36) Reynaldo, L. P.; Vologodskii, A. V; Neri, B. P.; Lyamichev, V. I. *J. Mol. Biol.* **2000**, *297*, 511−520.
(37) Wallace, D. C.; Brown, M. D.; Lott, M. T. *Gene* **1999**, *238*, 211−230.
(38) Coble, M. D.; Loreille, O. M.; Wadhams, M. J.; Edson, S. M.; Maynard, K.; Carna, E.; Berger, C.; Berger, B.; Falsetti, A. B.; Gill, P.; Niedersta, H.; Parson, W.; Finelli, L. N. *PLoS One* **2009**, *4*, e4838.
(39) Zuker, M. *Nucleic Acids Res.* **2003**, *31*, 3406−3415.

- D.A. Khodakov, **A.S. Khodakova**, A. Linacre, A.V. Ellis. Amelogenin Locus Typing Using Toehold-Assisted Fluorescent DNA Melting Analysis. **Forensic Science International: Genetics, 2013**, 4(1), 119-120.

Forensic Science International: Genetics Supplement Series 4 (2013) e119–e120

Contents lists available at ScienceDirect

Forensic Science International: Genetics Supplement Series

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

# Amelogenin locus typing using toehold-assisted fluorescent DNA melting analysis

Dmitriy A. Khodakov [a,*], Anastasia S. Khodakova [b], Adrian Linacre [b], Amanda V. Ellis [a,**]

[a] Flinders Centre for Nanoscale Science and Technology, Flinders University, Sturt Road, Bedford Park, Adelaide, South Australia, Australia
[b] School of Biological Sciences, Flinders University, Sturt Road, Bedford Park, Adelaide, South Australia, Australia

ARTICLE INFO

*Article history:*
Received 29 August 2013
Accepted 2 October 2013

*Keywords:*
DNA genotyping
DNA nanotechnology
Toehold displacement
Melting curve analysis
Amelogenin

ABSTRACT

The amelogenin gene is the locus of choice for gender identification in forensic science. Here we report on the use of fluorescent DNA melting curve analysis to genotype the amelogenin locus by means of a toehold-assisted DNA strand displacement reaction. The shape of the curves, or "polarity" of the melting peaks, allowed for visual discrimination between male and female DNA samples.



## 1. Introduction

Current DNA human profiling in forensic science uses polymerase chain reaction (PCR) based amplification of short tandem repeat (STR) loci along with the amelogenin gender marker gene. However, there is much scope for the development of faster methods of genetic testing which do not compromise on precision or reproducibility.

Recently, non-enzymatic DNA strand displacement reactions, which involve the formation of a so-called "toehold" structure, have gained acceptance in the fields of DNA nanotechnology, DNA computation and bio-sensing (for review see [1]). The toehold-mediated strand displacement process is a reaction between a double-stranded DNA duplex bearing a single-stranded overhang and a single-stranded oligonucleotide (displacing sequence) which is complementary to that of the longest strand of the duplex. A "zipper" mode, otherwise known as a consecutive "base-by-base" uni-directional re-hybridization mechanism, allows for the highly efficient and selective discrimination of the reacting DNA molecules even at the level of single nucleotide polymorphisms [2].

Previously we have demonstrated [3] a new approach for the DNA genotyping of real-life DNA samples in which a PCR product was generated from isolated genomic DNA by amplification with deoxyuracil modified forward and fluorescent dye (FAM) labelled reverse primers. Treatment of the obtained PCR product with uracil-DNA glycosylase (UDG) resulted in the formation of a PCR product with a single-stranded overhang (toehold-PCR product), suitable for toehold-mediated strand displacement reactions. Subsequently, the toehold-PCR product was reacted with a chemically synthesized single-stranded discriminating oligonucleotide (displacing sequence) labelled with the fluorescent quencher (TAMRA). By using the Förster resonance energy transfer (FRET) between the FAM and TAMARA fluorescent dyes, the kinetics of the displacement reaction was monitored in real-time. To this end the amelogenin gender marker locus was successfully genotyped as a model system. Furthermore, the specificity of the reaction was shown to be very high within a broad range of the reaction temperatures from 30 °C to 60 °C.

Here, instead of real-time monitoring the kinetics of the displacement reaction, we use a different approach based on fluorescent DNA melting analysis to assess the toehold-mediated DNA strand displacement products. Importantly, we show that this melting curve analysis can successfully discriminate between male and female amelogenin loci.

* Corresponding author. Tel.: +61 8 8201 2684; fax: +61 8 8201 2905.
** Corresponding author. Tel.: +61 8 8201 3104; fax: +61 8 8201 2905.
*E-mail addresses:* dmitriy.khodakov@flinders.edu.au (D.A. Khodakov), amanda.ellis@flinders.edu.au (A.V. Ellis).

1875-1768/$ – see front matter © 2013 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.fsigss.2013.10.061

## 2. Materials and methods

*Human genomic DNA* was extracted from the authors' own blood using a QIAamp DNA blood mini kit (Qiagen, Germany). PCR was performed using a HotStar Taq Master Mix kit (Qiagen, Germany). The forward and reverse primers with the following sequences 5′-d(CCC**U**GGGC**U**CTGTAAAGAATAGTG)-3′ and 5-FAM-d(ATCA-GAGCTTAAACTG GGAAGCT)-3′ (IDT-DNA, USA), respectively, were used at a final concentration of 0.2 μM each. Directly after the PCR, 2.5 U of the uracil-DNA glycosylase (NEB, USA) was added to the PCR solution and incubated at room temperature for 5 min. Then the mixture was heated to 95 °C for 5 min, cooled and purified using a Qiaquick PCR purification kit (Qiagen, Germany).

*Melting curve analysis* of the obtained toehold-PCR products (2pMole) (namely: a female toehold-PCR product containing amplicons from two X chromosomes (2X); and a male toehold-PCR product containing amplicons from one X and one Y chromosomes (XY)) was performed in the presence of 10× excess of the displacing oligonucleotides with following sequences: X*i* 5′-CCCTGGGCTCTGTAAAGAATAGTGTGTTGATTCTTTATCCCAGATGTT-TCTCAAGTGGTCCTGATTTTACA GTTCCTACCACCAGCTTCCCAGTT-TAAGCTCTGAT-TAMRA-3′ and Y*i* 5′-CCCTGGGCTCTGTAAAGAA-TAGTGGG TGGATTCTTCATCCCAAATAAAGTGGTTTCTCAAGTGGTC-CCAATTTTACAGTTCCTACCATCAGCTTCCCAGTTTAAGCTCTGAT-TAMRA-3′ (Eurogentec, Belgium). The melting curves were recorded using a RotorGene 3000 real-time PCR thermocycler.

## 3. Results and discussion

A 10× excess of the TAMRA labelled displacing sequences was added to the FAM labelled amelogenin toehold-PCR products (male and female). The mixture was then heated to 95 °C, at 0.1 °C/s and the change of the fluorescent intensity was recorded over a temperature range from 60 to 95 °C. As was shown previously [3] the strand displacement reaction proceeded rapidly at the elevated temperatures. Thus we presumed that the displacement reaction had terminated before the melting analysis initiated. Melting curves for all four possible combinations of two toehold-PCR products (male and female) and two displacing sequences (X*i* and Y*i* which were fully complementary to the longer strands of toehold-PCR products produced from the X and Y chromosomes, respectively) were obtained and plotted as a positive derivative of fluorescence with respect to temperature ($dF/dT$) against the temperature ($T$) (Fig. 1). We observed that if any displacement reaction between the displacing sequence and the toehold-PCR product occurred, this resulted in a "positive" melting peak located above $y = 0$ in Fig. 1. For example this was the case for the reaction between the female toehold-PCR (2X) product and the X*i* displacing sequence (Fig. 1, black hollow circles). As expected we also observed "positive" melting peaks for the reactions between the male toehold-PCR product (XY) and both X*i* (Fig. 1, red solid triangles) and Y*i* (Fig. 1, blue solid triangles) displacing sequences. The area under the curves for the male sample DNA was approximately half that of the female due to the fact that there was half the amount of the amplicons in the male toehold-PCR product reacting with either the displacing sequences. That is, two X's from the female toehold-PCR product react with two Xi's, and only one X or Y from the male toehold-PCR product reacts with one X*i* or one Y*i*, respectively).

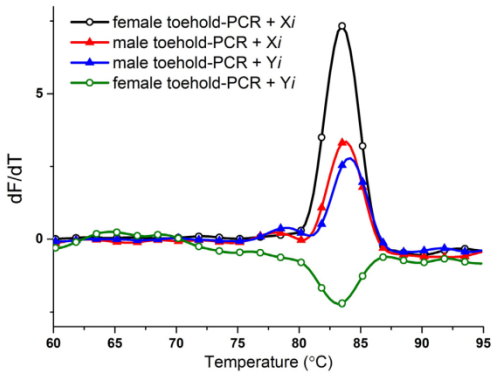



**Fig. 1.** Melting curve analysis of the toehold-mediated DNA strand displacement products of the amelogenin locus.

The reaction between the female toehold-PCR product (2X) and the Y*i* displacing sequence resulted in a "negative" melting peak located under the $y = 0$ (Fig. 1, green hollow circles). This "negative" peak indicates that no reaction of strand displacement occurred. This negative peak can be attributed to the quenched state of the FAM fluorophore on the unhybridized single-stranded oligonucleotide. This is characterized by the low mobility of the fluorescent dye formed during the melting along with the decrease in the fluorophore quantum yield [4].

## 4. Conclusion

We show that male and female amelogenin toehold-PCR products can be successfully discriminated based on the "polarity" of their melting peaks obtained after the strand displacement reaction with the Y*i* displacing sequence and regardless the peaks' melting temperatures. This opens up opportunities for simple, rapid and user friendly genetic sex discrimination in the forensic science.

## Role of funding

Funding was provided by the Department of Justice, South Australia.

## Conflict of interest

The authors declare no competing interest.

## References

[1] D.Y. Zhang, G. Seelig, Dynamic DNA nanotechnology using strand-displacement reactions, Nature Chemistry 3 (2011) 103–113.
[2] D.Y. Zhang, S.X. Chen, P. Yin, Optimizing the specificity of nucleic acid hybridization, Nature Chemistry 4 (2012) 208–214.
[3] D.A. Khodakov, A.S. Khodakova, A. Linacre, A.V. Ellis, Toehold-mediated nonenzymatic DNA strand displacement as a platform for DNA genotyping, Journal of the American Chemical Society 135 (2013) 5612–5619.
[4] I. Nazarenko, R. Pires, B. Lowe, M. Obaidy, A. Rashtchian, Effect of primary and secondary structure of oligodeoxyribonucleotides on the fluorescent properties of conjugated dyes, Nucleic Acids Research 30 (2002) 2089–2195.

# References


Aanderud, Z.T. et al., 2013. Sensitivity of soil respiration and microbial communities to altered snowfall. Soil Biology and Biochemistry, 57, pp.217–227.

Aitken, C.G.G. & Taroni, F., 2004. Statistics and the evaluation of forensic evidence for forensic scientis, Wiley.

Alaeddini, R., 2012. Forensic implications of PCR inhibition-A review. Forensic Science International Genetics, 6(3), pp.297–305.

Ali, T., Spreeuwers, L. & Veldhuis, R., 2011. Towards automatic forensic face recognition. Informatics Engineering and Information. pp.47–55.

Altschul, S., Gish, W. & Miller, W., 1990. Basic local alignment search tool. Journal of Molecular Biology, 215, pp.403–410.

Amorim, J.H. et al., 2012. A simple boiling-based DNA extraction for RAPD profiling of landfarm soil to provide representative metagenomic content. Genetics and Molecular Research : GMR, 11(1), pp.182–9.

Anderson, I.C. & Cairney, J.W.G., 2004. Diversity and ecology of soil fungal communities: increased understanding through the application of molecular techniques. Environmental Microbiology, 6(8), pp.769–79.

Anderson, M. & Willis, T., 2003. Canonical analysis of principal coordinates: a useful method of constrained ordination for ecology. Ecology, 84(2), pp.511–525.

Anderson, M.J., Gorley, R.N, Clarke, C.K., 2008. PERMANOVA+ for PRIMER: Guide to Software and Statistical Methods. Primer-E: Plymouth, UK.

Aquila, I. et al., 2014. The role of forensic botany in crime scene investigation: case report and review of literature. Journal of Forensic Sciences, 59(3), pp.820–824.

Ashayeri-Panah, M., Eftekhar, F. & Feizabadi, M.M., 2012. Development of an optimized random amplified polymorphic DNA protocol for fingerprinting of Klebsiella pneumoniae. Letters in Applied Microbiology, 54(4), pp.272–279.

Atienzar, F. & Jha, A., 2006. The random amplified polymorphic DNA (RAPD) assay and related techniques applied to genotoxicity and carcinogenesis studies: a critical review. Mutation Research, 613(2-3), pp.76–102.

Azad, A.K., Coote J.G., Parton R., 1991. An improved method for rapid purification of covalently closed circular plasmid DNA over a wide size range. Letters in Applied Microbiology, 14(6), pp.250-254.

Ballantyne, K.N., Poy, A.L. & van Oorschot, R. a. H., 2013. Environmental DNA monitoring: beware of the transition to more sensitive typing methodologies. Australian Journal of Forensic Sciences, 45(3), pp.323–340.

Bates, S.T. et al., 2011. Examining the global distribution of dominant archaeal populations in soil. The ISME Journal, 5(5), pp.908–917.

Bentley. D.R., et al., 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456, pp. 53-59.

Berg, M., 2013. Patterns of biodiversity at fine and small spatial scales. Soil Ecology and Ecosystem Services, pp.136–152.

Bergslien, E., 2012. An Introduction to Forensic Geoscience, Wiley-Blackwell.

Binga, E.K., Lasken, R.S. & Neufeld, J.D., 2008. Something from (almost) nothing: the impact of multiple displacement amplification on microbial ecology. The ISME Journal, 2(3), pp.233–241.

Biswas, A., Ranjan, D. & Zubair, M., Parallelization of MIRA Whole Genome and EST Sequence Assembler. Available at: http://www.cs.odu.edu/~zubair/papers/HPiC_Graph_2011_OCT8.pdf.

Bonetti, J. & Quarino, L., 2014. Comparative forensic soil analysis of New Jersey state parks using a combination of simple techniques with multivariate statistics. Journal of Forensic Sciences, 59(3), pp.627–636.

Boom R., Sol C.J., Salimans M.M., Jansen C.L., Wertheim-van Dillen P.M., van der Noordaa J., 1990. Rapid and simple method for purification of nucleic acids. Journal of Clinical Microbiology. 28(3); pp.495-503.

Boom R., Sol C.J., Heijtink R., Wertheim-van Dillen P.M., van der Noordaa J., 1991. Rapid purification of hepatitis B virus DNA from serum. Journal of Clinical Microbiology, 29(9); pp.1804-1811.

Di Bonito, R. et al., 2013. Characterization by length heterogeneity (LH)-PCR of a hydrogen-producing community obtained in dark fermentation using coastal lake sediment as an inoculum. Energy, Sustainability and Society, 3(1), p.3.

Brady, N.C. & Weil, R.R., 2008. The nature and properties of soils, Prentice Hall.

Brown, A.G., Smith, A. & Elmhurst, O., 2002. The combined use of pollen and soil analyses in a search and subsequent murder investigation. Journal of Forensic Sciences, 47(3), pp.614–618.

Buchan, A., Crombie, B. & Alexandre, G.M., 2010. Temporal dynamics and genetic diversity of chemotactic-competent microbial populations in the rhizosphere. Environmental Microbiology, 12(12), pp.3171–3184.

Budowle, B. et al., 2014. Validation of high throughput sequencing and microbial forensics applications. Investigative Genetics, 5(1), p.9.

Buermans, H. P. J., den Dunnen, J.T., 2014. Next generation sequencing technology: Advances and applications. Biochimica et Biophysica Acta, 1842, pp. 1932-1941.

Bull, P. A. et al., 2006. The transfer and persistence of trace particulates: experimental studies using clothing fabrics. Science & Justice : Journal of the Forensic Science Society, 46(3), pp.185–195.

Caetano-anollds, G., Bassam, B.J. & Gresshoff, P.M., 1991. DNA Amplification Fingerprinting: A Strategy for Genome Analysis. Plant Molecular Biology Reporter, 9(4), pp.294–307.

Caetano-Anolles, G., 1993. Amplifying DNA with arbitrary oligonucleotide primers. Genome Research, 3(2), pp.85–94.

Caetano-Anolles, G., 1994. MAAP: a versatile and universal tool for genome analysis. Plant Molecular Biology, pp.1011–1026.

Caporaso, J. & Lauber, C., 2011. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proceedings of the National Academy of Sciences of the United States of America, 108 Suppl, pp.4516–4522.

Caporaso, J.G. et al., 2010. QIIME allows analysis of high-throughput community sequencing data. Nature Methods, 7, pp.335–336.

Carlos, C., Castro, D.B.A. & Ottoboni, L.M.M., 2014. Comparative Metagenomic Analysis of Coral Microbial Communities Using a Reference-Independent Approach. PloS ONE, 9(11), p.e111626.

Cengiz, S. et al., 2004. SEM-EDS analysis and discrimination of forensic soil. Forensic Science International, 141(1), pp.33–37.

Champlot, S. et al., 2010. An efficient multistrategy DNA decontamination procedure of PCR reagents for hypersensitive PCR applications. PloS ONE, 5(9), p.e13042.

Chen, W. et al., 2013. A comparison of methods for clustering 16S rRNA sequences into OTUs. PloS ONE, 8(8), p.e70837.

Christopher M. Triggs, John S. Buckleton, S.J.W., 2004. Forensic DNA Evidence Interpretation, CRC Press.

Clarke, K., 1993. Non-parametric multivariate analyses of changes in community structure. Australian Journal of Ecology, (18), pp.117–143.

Clarke, K. & Gorley, R., 2006. PRIMER V6: User Manual/Tutorial. Primer-E: Plymouth, UK.

Clarke, K. & Green, R., 1988. Statistical design and analysis for a“ biological effects” study. Marine Ecology Progress Series, (46), pp. 213-226.

Clarke, K. & Warwick, R., 2001. Change in marine communities: an approach to statistical analysis and interpretation, PRIMER-E: Plymouth, UK.

Clarke, K.R., Somerfield, P.J. & Chapman, M.G., 2006. On resemblance measures for ecological studies, including taxonomic dissimilarities and a zero-adjusted Bray–Curtis coefficient for denuded assemblages. Journal of Experimental Marine Biology and Ecology, 330(1), pp.55–80.

Cock, P.J. et al., 2010. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Research, 38(6), pp.1767–1771.

Cole, J.R. et al., 2014. Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Research, (42), pp.D633–642.

Coleman, D., 2001. Soil biota, soil systems and processes. Encyclopedia of Biodiversity, (5), Academic Press.

Concheri, G. et al., 2011. Chemical elemental distribution and soil DNA fingerprints provide the critical evidence in murder case investigation. PloS ONE, 6(6), p.e20222.

Cox, R.J. et al., 2000. The forensic analysis of soil organic by FTIR. Forensic Science International, 108(2), pp.107–116.

Coyle, H.M., 2008. Nonhuman DNA Typing: Theory and Casework Applications, CRC Press Taylor&Francis Group.

Dabrowski, W. et al., 2003. Optimisation of AP-PCR fingerprinting discriminatory power for clinical isolates of Pseudomonas aeruginosa. FEMS Microbiology Letters, 218(1), pp.51–27.

Daniel, R. et al., 2015. A SNaPshot of next generation sequencing for forensic SNP analysis. Forensic Science International Genetics, (14), pp.50–60.

Dawson, L. A. & Hillier, S., 2010. Measurement of soil characteristics for forensic applications. Surface and Interface Analysis, 42(5), pp.363–377.

Delmont, T.O. et al., 2011. Accessing the soil metagenome for studies of microbial diversity. Applied and Environmental Microbiology, 77(4), pp.1315–1324.

Delmont, T.O. et al., 2012. Structure, fluctuation and magnitude of a natural grassland soil metagenome. The ISME Journal, 6(9), pp.1677–1687.

DeSantis, T.Z. et al., 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Applied and Environmental Microbiology, 72(7), pp.5069–5072.

DeSantis, T.Z. et al., 2007. High-density universal 16S rRNA microarray analysis reveals broader diversity than typical clone library when sampling the environment. Microbial Ecology, 53(3), pp.371–383.

Dineen, S.M. et al., 2010. An evaluation of commercial DNA extraction kits for the isolation of bacterial spore DNA from soil. Journal of Applied Microbiology, 109(6), pp.1886–1896.

Dong, D. et al., 2006. Removal of humic substances from soil DNA using aluminium sulfate. Journal of Microbiological Methods, 66(2), pp.217–222.

Dutilh, B.E. et al., 2012. Reference-independent comparative metagenomics using cross-assembly: crAss. Bioinformatics (Oxford, England), 28(24), pp.3225–3231.

Edgar, R.C., 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics (Oxford, England), 26(19), pp.2460–2471.

Edmond, G. & Mercer, D., 2004. Daubert and the Exclusionary Ethos: The Convergence of Corporate and Judicial Attitudes towards the Admissibility of Expert Evidence in Tort Litigation*. Law & Policy, 26(2), pp.232-257.

Epp, L.S. et al., 2012. New environmental metabarcodes for analysing soil DNA: potential for studying past and present ecosystems. Molecular Eecology, 21(8), pp.1821–1833.

Esteban, G.F. et al., 2006. Soil protozoa—An intensive study of population dynamics and community structure in an upland grassland. Applied Soil Ecology, 33(2), pp.137–151.

Evett, I. et al., 2000. The impact of the principles of evidence interpretation on the structure and content of statements. Science & Justice, 40(4), pp.233–239.

Ewing, B. et al., 1998. Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. Genome Research, pp.175–185.

Ewing, B. & Green, P., 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Research, (206), pp.186–194.

Fatima, F., Chaudhary, I. & Ali, J., 2011. Microbial DNA extraction from soil by different methods and its PCR amplification. Biochemical Cellular Archive, 11(1), pp. 2011-2017.

Feinstein, L.M., Sul, W.J. & Blackwood, C.B., 2009. Assessment of bias associated with incomplete extraction of microbial DNA from soil. Applied and Environmental Microbiology, 75(16), pp.5428–5433.

Fierer, N. et al., 2012. Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. Proceedings of the National Academy of Sciences of the United States of America, 109(52), pp.21390–21395.

Fierer, N. et al., 2007. Metagenomic and small-subunit rRNA analyses reveal the genetic diversity of bacteria, archaea, fungi, and viruses in soil. Applied and Environmental Microbiology, 73(21), pp.7059–7066.

Fitzpatrick, R., 2009. Soil: Forensic Analysis. In A. Jamieson & A. Moenssens, eds. Wiley Encyclopedia of Forensic Science. The Atrium, Southern Gate, Chichester, West Sussex, United Kingdom: John Wiley&Sons Ltd, pp. 2377–2388.

Fitzpatrick, R.W., 2013. Soil : Forensic Analysis. In Wiley Encyclopedia of Forensic Science. Wiley.

Fitzpatrick, R.W. & Raven, M.D., 2013. Guidelines for Conducting Criminal and Environmental Soil Forensic Investigations.

Fitzpatrick, R.W. & Raven, M.D., 2012. How Pedology and Mineralogy Helped Solve a Double Murder Case: Using Forensics to Inspire Future Generations of Soil Scientists. Soil Horizons, 53(5), p.14.

Fordyce, S.L. et al., 2015. Second-generation sequencing of forensic STRs using the Ion TorrentTM HID STR 10-plex and the Ion PGMTM. Forensic Science International Genetics, 14, pp.132–140.

Franklin, R.B., Taylor, D.R. & Mills, A.L., 1999. Characterization of microbial communities using randomly amplified polymorphic DNA (RAPD). Journal of Microbiological Methods, 35(3), pp.225–235.

French, J.C. et al., 2012. Multiple transfers of particulates and their dissemination within contact networks. Science and Justice, 52, pp.33–41.

Fuhrman, J. A, 2012. Metagenomics and its connection to microbial community organization. F1000 Biology Reports, (4), p.15.

Giampaoli, S. et al., 2014. The environmental biological signature: NGS profiling for forensic comparison of soils. Forensic Science International, 240, pp.41–47.

Gilbert, J. a et al., 2010. The taxonomic and functional diversity of microbes at a temperate coastal site: a “multi-omic” study of seasonal and diel temporal variation. PloS ONE, 5(11), p.e15545.

Gilbert, R.O. & Pulsipher, B.A., 2005. Role of Sampling Designs in Obtaining Representative Data. Environmental Forensics, 6, pp.27–33.

Gillan, R. et al., 1995. Comparison of Cannabis sativa by Random Amplification of Polymorphic DNA (RAPD) and HPLC of cannabinoids: a preliminary study. Science & Justice, 35(3), pp.1–9.

Giri, B., Giang, P. & Kumari, R., 2005. Microbial diversity in soils. In Microorganisms in soils: Roles in Genesis and Functions, Soil Biology. pp. 19–55.

Gonzalez-Rodriguez, J., 2006. Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition. Computer Speech & Language, (20), pp. 331-355.

Grasis, J. a et al., 2014. Species-specific viromes in the ancestral holobiont hydra. PloS ONE, 9(10), p.e109952.

Haas, B.J. et al., 2011. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. Genome research, 21(3), pp.494–504.

Hajibabaei, M., 2012. The golden age of DNA metasystematics. Trends in Genetics : TIG, 28(11), pp.535–537.

Handelsman, J., 2004. Metagenomics: application of genomics to uncultured microorganisms. Microbiology and Molecular Biology Reviews, 68(4), pp.669–685.

Handelsman, J., Rondon, M. & Brady, S., 1998. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chemistry & Biology, (5), pp. 245-249.

Håvelsrud, O.E. et al., 2012. Metagenomic and geochemical characterization of pockmarked sediments overlaying the Troll petroleum reservoir in the North Sea. BMC Microbiology, (12), pp.203-207.

Heath, L.E. & Saunders, V. A, 2006. Assessing the potential of bacterial DNA profiling for forensic soil comparisons. Journal of Forensic Sciences, 51(5), pp.1062–1068.

Heinrich, E.W., 1965. Microscopic Identification of Minerals, McGraw-Hill, New York, NY.

Helton, R.R. & Wommack, K.E., 2009. Seasonal dynamics and metagenomic characterization of estuarine viriobenthos assemblages by randomly amplified polymorphic DNA PCR. Applied and Environmental Microbiology, 75(8), pp.2259–2265.

Heo, J. & Kim, S., 2013. Cloning the Pfu DNA polymerase from DNA contaminants in preparations of commercial Pfu DNA polymerase. African Journal of Microbiology Research, 7(9), pp.745–750.

Horrocks, M., Coulson, S. & Walsh, K., 1999. Forensic palynology: variation in the pollen content of soil on shoes and in shoeprints in soil. Journal of Forensic Sciences, (6), pp.119–122.

Horswell, J. et al., 2002. Forensic comparison of soils by bacterial community DNA profiling. Journal of Forensic Sciences, 47(2), pp.350–353.

Houck, M., 2004. Trace evidence analysis. More cases in mute witnesses, Elsevier Academic Press.

Howe, A.C., Jansson, J.K., Malfatti, S. a, et al., 2014. Tackling soil diversity with the assembly of large, complex metagenomes. Proceedings of the National Academy of Sciences of the United States of America, 111(13), pp.4904–4909.

Hubert, F. et al., 2009. Advances in characterization of soil clay mineralogy using X-ray diffraction: from decomposition to profile fitting. European Journal of Soil Science, 60(6), pp.1093–1105.

Hur, I. & Chun, J., 2004. A Method for Comparing Multiple Bacterial Community Structures from 16S rDNA Clone Library Sequences. The Journal of Microbiology, 42(1), pp.9–13.

Iulia, L. & Bianca, I., 2013. The evidence of contaminant bacterial DNA in several commercial Taq polymerases. Romanian Biotechnological Letters, 18(1), pp.8007–8012.

Iyengar, A. & Hadi, S., 2014. Use of non-human DNA analysis in forensic science: a mini review. Medicine, Science and The Law, 54(1), pp.41–50.

Jackson, C.R. et al., 1997. A simple, efficient method for the separation of humic substances and DNA from environmental samples. Applied and Environmental Microbiology, 63(12), pp.4993–4995.

Jagadish, V., Robertson, J. & Gibbs, A., 1996. RAPD analysis distinguishes Cannabis sativa samples from different sources. Forensic Science International, 79, pp.113–121.

Jain, A., Murty, M. & Flynn, P., 1999. Data clustering: a review. ACM computing surveys (CSUR), 31(3), pp.264-323.

Janssen, P., 2006. Identifying the dominant soil bacterial taxa in libraries of 16S rRNA and 16S rRNA genes. Applied and Environmental Microbiology, 72(3), pp. 1719-1728.

Jeffries, T.C. et al., 2011. Substrate type determines metagenomic profiles from diverse chemical habitats. PloS ONE, 6(9), p.e25173.

Jenkins, S.N. et al., 2010. Taxon-specific responses of soil bacteria to the addition of low level C inputs. Soil Biology and Biochemistry, 42(9), pp.1624–1631.

Jhang, T. & Shasany, A.K., 2012. Plant DNA Fingerprinting and Barcoding N. J. Sucher, J. R. Hennell, & M. C. Carles, eds., Totowa, NJ: Humana Press.

Jiang, B. et al., 2012. Comparison of metagenomic samples using sequence signatures. BMC genomics, 13(1), pp.730-742.

Kennedy, A. & Stubbs, T., 2006. Soil microbial communities as indicators of soil health. Annals of Arid zone, 45, pp.287–308.

Kent, W.J., 2002. BLAT — The BLAST -Like Alignment Tool. Genome Research, (12) , pp.656–664.

Khandka, D., Tuna, M. & Tal, M., 1997. Variability in the pattern of random amplified polymorphic DNA. Electrophoresis, (18), pp.2852–2856.

Khodakov, D. et al., 2008. An oligonucleotide microarray for multiplex real-time PCR identification of HIV-1, HBV, and HCV. BioTechniques, 44(2), pp.241–248.

Khodakova, A.S. et al., 2013. Forensic analysis of soils using single arbitrarily primed amplification and high throughput sequencing. Forensic Science International: Genetics Supplement Series, 4(1), pp.e39–e40.

Kirk, J.L. et al., 2004. Methods of studying soil microbial diversity. Journal of Microbiological Methods, 58(2), pp.169–88.

Klammer, S., Mondini, C. & Insam, H., 2005. Microbial community fingerprints of composts stored under different conditions. Annals of Microbiology, 55(4), pp.299–305.

Klaschik, S. et al., 2002. Comparison of different decontamination methods for reagents to detect low concentrations of bacterial 16S DNA by real-time-PCR. Molecular Biotechnology, 22(3), pp.231–242.

Knauth, S., Schmidt, H. & Tippkötter, R., 2013. Comparison of commercial kits for the extraction of DNA from paddy soils. Letters in Applied Microbiology, 56(3), pp.222–8.

Lauber, C.L. et al., 2010. Effect of storage conditions on the assessment of bacterial community structure in soil and human-associated samples. FEMS Microbiology Letters, 307(1), pp.80–86.

Lauber, C.L. et al., 2013. Temporal variability in soil microbial communities across land-use types. The ISME Journal, 7(8), pp.1641–1650.

Leckie, S., 2005. Methods of microbial community profiling and their application to forest soils. Forest Ecology and Management, (220) pp. 88-106 .

Lerner, A. et al., 2006. Can denaturing gradient gel electrophoresis (DGGE) analysis of amplified 16s rDNA of soil bacterial populations be used in forensic investigations? Soil Biology and Biochemistry, 38, pp.1188–1192.

Lilje, L. et al., 2013. Soil sample metagenome NGS data management for forensic investigation. Forensic Science International: Genetics Supplement Series, 4(1), pp.e35–e36..

Liu, W.T. et al., 1997. Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA. Applied and Environmental Microbiology, 63(11), pp.4516–4522.

Logares, R. et al., 2012. Environmental microbiology through the lens of high-throughput DNA sequencing: synopsis of current platforms and bioinformatics approaches. Journal of Microbiological Methods, 91(1), pp.106–113.

Loman, N.J. et al., 2012. Performance comparison of benchtop high-throughput sequencing platforms. Nature Biotechnology, 30(5), pp.434–439.

Macdonald, C.A. et al., 2011. Discrimination of soils at regional and local levels using bacterial and fungal T-RFLP profiling. Journal of Forensic Sciences, 56(1), pp.61–69.

Macdonald, L.M. et al., 2008. Microbial DNA profiling by multiplex terminal restriction fragment length polymorphism for forensic comparison of soil and the influence of sample condition. Journal of Applied Microbiology, 105(3), pp.813–821.

Maillet, N. et al., 2012. Compareads: comparing huge metagenomic experiments. BMC Bioinformatics, 13 (Suppl 19), p.S10.

Mande, S.S., Mohammed, M.H. & Ghosh, T.S., 2012a. Classification of metagenomic sequences: methods and challenges. Briefings in Bioinformatics, 13(6), pp.669–681.

Margulies, M. et al., 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature, 437(7057), pp.376–380.

Markowitz, V.M. et al., 2012. IMG: the Integrated Microbial Genomes database and comparative analysis system. Nucleic Acids Research, (40), pp.D115–122.

Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.Journal, 17(1), p.10.

Mason, O.U. et al., 2014. Metagenomics reveals sediment microbial community response to Deepwater Horizon oil spill. The ISME Journal, pp.1–12.

McCauley, A., Jones, C. & Jacobsen, J., 2005. Basic soil properties. Montana State University Extention Service: Soil& Water, pp.1–12.

McCrone, W.C., Draftz, R.G. & Delly, J.G., 1967. The Particle Atlas. A photomicrographic reference for the microscopical identification of particulate substances, Ann Arbor Science Publishers, Inc.

McDonald, D. & Clemente, J., 2012. The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. GigaScience, (1), pp.1–6.

McWilliam, H. et al., 2013. Analysis Tool Web Services from the EMBL-EBI. Nucleic Acids Research, (41), pp.W597–600.

Menking, D. et al., 1999. Rapid cleanup of bacterila DNA from field samples. Conservation and Recycling, 27, pp.179–186.

Mennerat, A. & Sheldon, B.C., 2014. How to Deal with PCR Contamination in Molecular Microbial Ecology. Microbial Ecology, 68(4), pp. 834-841.

Meyer, F. et al., 2008. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics, 9, p.386.

Meyers, M.S. & Foran, D.R., 2008. Spatial and temporal influences on bacterial profiling of forensic soil samples. Journal of Forensic Sciences, 53(3), pp.652–660.

Miao, T. et al., 2014. A method suitable for DNA extraction from humus-rich soil. Biotechnology Letters, 36(11), pp. 2223-2228.

Mildenhall, D.C., Wiltshire, P.E.J. & Bryant, V.M., 2006. Forensic palynology: why do it and how it works. Forensic Science International, 163(3), pp.163–172.

Miller, D.N. et al., 1999. Evaluation and optimization of DNA extraction and purification procedures for soil and sediment samples. Applied and Environmental Microbiology, 65(11), pp.4715–4724.

Moreno, L.I. et al., 2006. Microbial metagenome profiling using amplicon length heterogeneity-polymerase chain reaction proves more effective than elemental analysis in discriminating soil specimens. Journal of Forensic Sciences, 51(6), pp.1315–1322.

Moreno, L.I. et al., 2011. The application of amplicon length heterogeneity PCR (LH-PCR) for monitoring the dynamics of soil microbial communities associated with cadaver decomposition. Journal of Microbiological Methods, 84(3), pp.388–393.

Morgan, R.M., Flynn, J., et al., 2014. Experimental forensic studies of the preservation of pollen in vehicle fires. Science & Justice : Journal of the Forensic Science Society, 54(2), pp.141–145.

Morgan, R.M. et al., 2006. The role of forensic geoscience in wildlife crime detection. Forensic Science International, 162(1-3), pp.152–162.

Morgan, R.M., Allen, E., et al., 2014. The spatial and temporal distribution of pollen in a room: forensic implications. Science & Justice : Journal of the Forensic Science Society, 54(1), pp.49–56.

Morgan, R.M. & Bull, P. a., 2006. Data Interpretation in Forensic Sediment and Soil Geochemistry. Environmental Forensics, 7(4), pp.325–334.

Morgan, R.M. & Bull, P. A., 2007. The philosophy, nature and practice of forensic sediment analysis. Progress in Physical Geography, 31(1), pp.43–58.

Morrisson, A., McColl, S. & Dawson, L., 2009. Characterisation and Discrimination of Urban Soils: Preliminary Results from The Soil Forensics University Network. Soil Forensics, pp.75–86.

Mühl, H. et al., 2010. Activity and DNA contamination of commercial polymerase chain reaction reagents for the universal 16S rDNA real-time polymerase chain reaction detection of bacterial pathogens in blood. Diagnostic Microbiology and Infectious Disease, 66(1), pp.41–49.

Murray, R., 2011. Evidence From The Earth: Forensic Geology and Criminal Investigation, Mountain Press.

Murray, R., 2004. Evidence From the Earth: Forensic Geology and Criminal Investigation, Mountain Press.

Murray, R. & Tedrow, J.C.F., 1991. Forensic geology, Prentice Hall, Englewood Cliffs.

Murray, R. & Tedrow, J.C.F., 1975. Forensic Geology: Earth Sciences and Criminal Investigation, Rutgers University Press, New York, NY.

Muyzer, G., 1999. Genetic fingerprinting of microbial communities – present status and future perspectives. Methods of microbial community analysis. Available at: http://socrates.acadiau.ca/isme/Symposium16/muyzer.pdf.

Muyzer, G., 1993. Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. Applied and Environmental Microbiology, 59(3), pp.695-701.

Muyzer, G. & Smalla, K., 1998. Application of denaturing gradient gel electrophoresis (DGGE) and temperature gradient gel electrophoresis (TGGE) in microbial ecology. Antonie van Leeuwenhoek, 73(1), pp.127–141.

Namiki, T. et al., 2012. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. Nucleic Acids Research, 40(20), p.e155.

Nemergut, D.R. et al., 2011. Global patterns in the biogeography of bacterial taxa. Environmental Microbiology, 13(1), pp.135–144.

Ning, J. et al., 2009. Different influences of DNA purity indices and quantity on PCR-based DGGE and functional gene microarray in soil microbial community study. Applied Microbiology and Biotechnology, 82(5), pp.983–993.

Olson, N.D. & Morrow, J.B., 2012. DNA extract characterization process for microbial detection methods development and validation. BMC Research Notes, 5(1), p.668-681.

Van Oorschot, R.A., Ballantyne, K.N. & Mitchell, R.J., 2010. Forensic trace DNA: a review. Investigative Genetics, 1(1), p.14-30.

Overbeek, R. et al., 2005. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Research, 33(17), pp.5691–5702.

Pace, N.R., 1997. A Molecular View of Microbial Diversity and the Biosphere. Science, 276(5313), pp.734–740.

Paraguison, R.C. et al., 2012. Improved RAPD-PCR for discriminating breeds of water buffalo. Biochemical Genetics, 50(7-8), pp.579–584.

Parks, D.H. & Beiko, R.G., 2010. Identifying biologically relevant differences between metagenomic communities. Bioinformatics (Oxford, England), 26(6), pp.715–721.

Pasternak, Z., Al-Ashhab, A. & Gatica, J., 2012. Optimization of molecular methods and statistical procedures for forensic fingerprinting of microbial soil communities. International Research Journal of Microbiology, 3(11), pp.363–372.

Peckham, H.E., et al., 2007. SOLiD sequencing and 2-base encoding. San Diego, CA: American Society of Human Genetics.

Pelt-Verkuil, E. van, Belkum, A. van & Hays, J.P., 2008. Principles and Technical Aspects of PCR Amplification, Springer.

Peng, Y. et al., 2011. Meta-IDBA: a de Novo assembler for metagenomic data. Bioinformatics (Oxford, England), 27(13), pp.194–101.

Pennisi, E., 2014. DNA Sequencers Still Waiting for the Nanopore Revolution. Science, 343(February), pp.829–830.

Petric, I. et al., 2011. Inter-laboratory evaluation of the ISO standard 11063 "Soil quality - Method to directly extract DNA from soil samples". Journal of Microbiological Methods, 84(3), pp.454–460.

Petrosino, J.F. et al., 2009. Metagenomic pyrosequencing and microbial identification. Clinical Chemistry, 55(5), pp.856–866.

Pirrie, D. et al., 2014. Soil forensics as a tool to test reported artefact find sites. Journal of Archaeological Science, 41, pp.461–473.

Pirrie, D., Donnelly, L. & Rollinson, G., 2013. Forensic geology at the International School Science Fair 2013. Geology Today, 29(6), pp.222–228.

Pirrie, D., Ruffell, A. & Dawson, L. A., 2013. Environmental and criminal geoforensics: an introduction. Geological Society, 384(1), pp.1–7.

Prakash, T. & Taylor, T.D., 2012. Functional assignment of metagenomic data: challenges and applications. Briefings in Bioinformatics, 13(6), pp.711–727.

Pye, K., 2007. Geological and Soil Evidence: Forensic Applications, CRC Press Taylor&Francis Group.

Pye, K. & Croft, D.J., 2004. Forensic Geoscience: Principles, Techniques and Applications., Geological Society, London, Special Publication.

Pyrek, K., 2007. Forensic science under siege. Elsevier Academic Press.

Quast, C. et al., 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Research, (41), pp.D590–596.

Rajendhran, J. & Gunasekaran, P., 2008. Strategies for accessing soil metagenome for desired applications. Biotechnology Advances, 26(6), pp.576–590.

Rastogi, G. & Sani, R.K., 2011. Microbes and Microbial Technology I. Ahmad, F. Ahmad, & J. Pichtel, eds., New York, NY: Springer New York.

Rawlins, B.G. et al., 2006. Potential and pitfalls in establishing the provenance of Earth-related samples in forensic investigations. Journal of Forensic Sciences, 51(4), pp.832–845.

Rincon-Florez, V., Carvalhais, L. & Schenk, P., 2013. Culture-Independent Molecular Tools for Soil and Rhizosphere Microbiology. Diversity, 5(3), pp.581–612.

Rissanen, A.J. et al., 2010. Storage of environmental samples for guaranteeing nucleic acid yields for molecular microbiological studies. Applied Microbiology and Biotechnology, 88(4), pp.977–984.

Ritz, K., Dawson, L. & Miller, D., 2009. Criminal and Environmental Soil Forensics, Springer.

Roesch, L.F.W. et al., 2007. Pyrosequencing enumerates and contrasts soil microbial diversity. The ISME Journal, 1(4), pp.283–290.

Roose-Amsaleg, C.., Garnier-Sillam, E. & Harry, M., 2001. Extraction and purification of microbial DNA from soil and sediment samples. Applied Soil Ecology, 18(1), pp.47–60.

Rosen, M.J. et al., 2012. Denoising PCR-amplified metagenome data. BMC Bioinformatics, 13, p.283-269.

Rubin, B.E.R. et al., 2013. Investigating the impact of storage conditions on microbial community composition in soil samples. PloS ONE, 8(7), p.e70460.

Rudi, K. & Larsen, F., 1997. Strain Characterization and Classification of Oxyphotobacteria in Clone Cultures on the Basis of 16S rRNA Sequences. Applied and Environmental Microbiology, 63(7), pp.2593-2600.

Ruffell, A. & McKinley, J., 2005. Forensic geoscience: applications of geology, geomorphology and geophysics to criminal investigations. Earth-Science Reviews, 69(3-4), pp.235–247.

Ruffell, A. & Wiltshire, P., 2004. Conjunctive use of quantitative and qualitative X-ray diffraction analysis of soils and rocks for forensic analysis. Forensic Science International, 145(1), pp.13–23.

Sagar, K. et al., 2014. Assessment of five soil DNA extraction methods and a rapid laboratory-developed method for quality soil DNA extraction for 16S rDNA-based amplification and library construction. Journal of Microbiological Methods, 97, pp.68–73.

Sanger, F., Nicklen, S., Coulson, A.R., 1977. DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America, 74(12), pp. 5463-5467.

Schloss, P.D. & Westcott, S.L., 2011. Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis. Applied and Environmental Microbiology, 77(10), pp.3219–3226.

Scott, K.R. et al., 2014. The transferability of diatoms to clothing and the methods appropriate for their collection and analysis in forensic geoscience. Forensic Science International, 241, pp.127–137.

Sensabaugh, G.F., 2009. Microbial Community Profiling for the Characterisation of Soil Evidence : Forensic Considerations. In K. Ritz, L. Dawson, & D. Miller, eds. Criminal and Environmental Soil Forensics. Springer, pp. 49–60.

Shade, A. et al., 2014. Conditionally rare taxa disproportionately contribute to temporal changes in microbial diversity. mBio, 5(4), pp.e01371–1314.

Shange, R. et al., 2013. Assessing the Diversity and Composition of Bacterial Communities across a Wetland, Transition, Upland Gradient in Macon County Alabama. Diversity, 5(3), pp.461–478.

Sharma, S., Sharma, K.K. & Kuhad, R.C., 2014. An efficient and economical method for extraction of DNA amenable to biotechnological manipulations, from diverse soils and sediments. Journal of Applied Microbiology, 116(4), pp.923–933.

Shendure, J. & Ji, H., 2008. Next-generation DNA sequencing. Nature Biotechnology, 26(10), pp.1135–1145.

Shokralla, S. et al., 2012. Next-generation sequencing technologies for environmental DNA research. Molecular Ecology, 21(8), pp.1794–1805.

Shoyama, Y., Kawachi, F. & Tanaka, H., 1998. Genetic and alkaloid analysis of Papaver species and their F1 hybrid by RAPD, HPLC and ELISA. Forensic Science International, (91), pp.207–217.

Simon, C. & Daniel, R., 2011. Metagenomic analyses: past and future trends. Applied and Environmental Microbiology, 77(4), pp.1153–1161.

Smith, R.J. et al., 2013. Determining the metabolic footprints of hydrocarbon degradation using multivariate analysis. PloS ONE, 8(11), p.e81910

Spangler, R., Goddard, N.L. & Thaler, D.S., 2009. Optimizing Taq polymerase concentration for improved signal-to-noise in the broad range detection of low abundance bacteria. PloS ONE, 4(9), p.e7010.

Spiegelman, D., Whissell, G. & Greer, C.W., 2005. A survey of the methods for the characterization of microbial consortia and communities. Canadian Journal of Microbiology, 51(5), pp.355–386.

Srinivasiah, S. et al., 2013. Direct assessment of viral diversity in soils by random PCR amplification of polymorphic DNA. Applied and Environmental Microbiology, 79(18), pp.5450–5457.

Stevenson, F.J., 1982. Humus Chemistry: Genesis, Composition, Reactions, Wiley.

Suenaga, H., 2012. Targeted metagenomics: a high-resolution metagenomics approach for specific gene clusters in complex microbial communities. Environmental Microbiology, 14(1), pp.13–22.

Taberlet, P. et al., 2012. Soil sampling and isolation of extracellular DNA from large amount of starting material suitable for metabarcoding studies. Molecular Ecology, 21, pp.1816–1820.

Takahashi, H. et al., 2014. Preparation of Phi29 DNA polymerase free of amplifiable DNA using ethidium monoazide, an ultraviolet-free light-emitting diode lamp and trehalose. PloS ONE, 9(2), p.e82624.

Taş, N. et al., 2014. Impact of fire on active layer and permafrost microbial communities and metagenomes in an upland Alaskan boreal forest. The ISME Journal, pp.1–16.

Terrat, S. et al., 2014. Meta-barcoded evaluation of the ISO standard 11063 DNA extraction procedure to characterize soil bacterial and fungal community diversity and composition. Microbial Biotechnology, (4), pp.12162-12148.

Terrat, S. et al., 2012. Molecular biomass and MetaTaxogenomic assessment of soil microbial communities as influenced by soil DNA extraction procedure. Microbial Biotechnology, 5(1), pp.135–41.

Torsvik, V. & Øvreås, L., 2002. Microbial diversity and function in soil: from genes to ecosystems. Current Opinion in Microbiology, pp.240–245.

Trindade-Silva, A.E. et al., 2012. Taxonomic and functional microbial signatures of the endemic marine sponge Arenosclera brasiliensis. PloS ONE, 7(7), p.e39905.

Tringe, S. & Rubin, E., 2005. Metagenomics: DNA sequencing of environmental samples. Nature Reviews Genetics, 6(11), pp.805-814.

Tringe, S.G. & Hugenholtz, P., 2008. A renaissance for the pioneering 16S rRNA gene. Current Opinion in Microbiology, 11(5), pp.442–446.

Tyler, K., Wang, G. & Tyler, S., 1997. Factors affecting reliability and reproducibility of amplification-based DNA fingerprinting of representative bacterial pathogens. Journal of Clinical Microbiology, 35(2), pp.339–346.

Tzeneva, V. a et al., 2009. Effect of soil sample preservation, compared to the effect of other environmental variables, on bacterial and eukaryotic diversity. Research in Microbiology, 160(2), pp.89–98.

Uroz, S. et al., 2013. Functional assays and metagenomic analyses reveals differences between the microbial communities inhabiting the soil horizons of a Norway spruce plantation. PloS ONE, 8(2), p.e55929.

Valverde, J.R. & Mellado, R.P., 2013. Analysis of metagenomic data containing high biodiversity levels. PloS ONE, 8(3), p.e58118.

Verma, K., 2013. Role of Diatoms in the World of Forensic Science. Journal of Forensic Research, 04(02), pp.2–5.

Vettori, C. et al., 1996. Amplification of bacterial DNA bound on clay minerals by the random amplified polymorphic DNA (RAPD) technique. FEMS Microbiology Ecology, 20(4), pp.251–260.

Vickery, M.C.. et al., 1998. Optimization of the arbitrarily-primed polymerase chain reaction (AP-PCR) for intra-species differentiation of Vibrio vulnificus. Journal of Microbiological Methods, 33(2), pp.181–189.

Vishnivetskaya, T. a et al., 2014. Commercial DNA extraction kits impact observed microbial community composition in permafrost samples. FEMS Microbiology Ecology, 87(1), pp.217–230.

Voříšková, J. & Brabcová, V., 2014. Seasonal dynamics of fungal communities in a temperate oak forest soil. New Phytologyst, (201) pp.269–278.

Wallenius, K. et al., 2010. Sample storage for soil enzyme activity and bacterial community profiles. Journal of Microbiological Methods, 81(1), pp.48–55.

Walsh, K. a J. & Horrocks, M., 2008. Palynology: its position in the field of forensic science. Journal of Forensic Sciences, 53(5), pp.1053–1060.

Wang, J. et al., 2013. Environmental bio-monitoring with high-throughput sequencing. Briefings in Bioinformatics, 14(5), pp.575–588.

Waters, J.M. et al., 2012. Arbitrary single primer amplification of trace DNA substrates yields sequence content profiles that are discriminatory and reproducible. Electrophoresis, 33(3), pp.492–498.

Welsh, J. & McClelland, M., 1990. Fingerprinting genomes using PCR with arbitrary primers. Nucleic Acids Research, 18(24), pp.7213–7218.

Wheeler, B. & Wilson, L.J., 2008. Practical Forensic Microscopy: A Laboratory Manual, Wiley.

Whitehouse, C. a & Hottel, H.E., 2007. Comparison of five commercial DNA extraction kits for the recovery of Francisella tularensis DNA from spiked soil samples. Molecular and Cellular Probes, 21(2), pp.92–96.

Wikström, P., Andersson, A.-C. & Forsman, M., 1999. Biomonitoring complex microbial communities using random amplified polymorphic DNA and principal component analysis. Soil Biology and Biochemistry, 30(2), pp.265–268.

Wilke, A. et al., 2014. MG-RAST Manual.

Wilke, A. et al., 2013. MG-RAST Technical report and manual for version 3.3. 6–rev. Available at: ftp://bastogne-2.mcs.anl.gov/data/manual/mg-rast-tech-report-v3_r1.pdf.

Williams, J.G. et al., 1990. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Research, 18(22), pp.6531–6535.

Williamson, K.E. et al., 2011. Optimizing the indirect extraction of prokaryotic DNA from soils. Soil Biology and Biochemistry, 43(4), pp.736–748.

Woods, B. et al., 2014. Soil examination for a forensic trace evidence laboratory-Part 1: Spectroscopic techniques. Forensic Science International, pp.1–8.

Wu, D. et al., 2009. A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. Nature, 462(7276), pp.1056–1060.

Xu, Z. et al., 2014. Bioinformatic approaches reveal metagenomic characterization of soil microbial community. PloS ONE, 9(4), p.e93445.

Yeates, G.W., 2003. Nematodes as soil indicators : functional and biodiversity aspects. Biology and Fertility of Soils, (37), pp.199–210.

Young, J.M., Weyrich, L.S. & Cooper, A., 2014. Forensic soil DNA analysis using high-throughput sequencing: A comparison of four molecular markers. Forensic Science International. Genetics, (13), pp.176–184.

Zadora, G. et al., 2013. Statistical Analysis in Forensic Science: Evidential Values of Multivariate Physicochemical Data, Wiley.

Zala, K., 2007. Dirty science: soil forensics digs into new techniques. Science, 318, pp.386–387.

Zarraonaindia, I., Smith, D.P. & Gilbert, J. A, 2013. Beyond the genome: community-level analysis of the microbial world. Biology & Philosophy, 28(2), pp.261–282.

Zhao, F. & Xu, K., 2012. Efficiency of DNA extraction methods on the evaluation of soil microeukaryotic diversity. Acta Ecologica Sinica, 32(4), pp.209–214.