Gene Expression Biomarkers for Colorectal Neoplasia

by

L. C. LaPointe

B.Sc. (Florida State University) 1991

Department of Medicine

FLINDERS UNIVERSITY OF SOUTH AUSTRALIA

Committee in charge:

Prof. Graeme P. Young, Chair

Dr. Robert A. Dunne
Dr. Peter L. Molloy 2008

Contents

Gene Expression Biomarkers for Colorectal Neoplasia

L. C. LaPointe

Flinders University of South Australia

Department of Medicine

Prof. Graeme P. Young

The aim of this research was to assemble sufficient experimental evidence about candidate gene transcript expression changes between non-neoplastic and neoplastic colorectal tissues to justify future assay development involving promising leads. To achieve this aim, this thesis explores the hypothesis that gene expression-based biomarkers can be used to accurately discriminate colorectal neoplastic tissues from non-neoplastic controls.

This hypothesis was tested by first analysing multiple, large, quality controlled data sets comprising gene expression measurements across colorectal phenotypes to discover potential biomarkers. Candidate biomarkers were then subjected to validation testing using a custom-design oligonucleotide microarray applied to independently derived clinical specimens. A number of novel conclusions are reached based on these data. The most important conclusion is that a defined subset of genes expressed in the colorectal mucosa are reliably differentially expressed in neoplastic tissues. In particular, the apparently high prediction accuracy achieved for single gene transcripts to discriminate hundreds of neoplastic and non-neoplastic tissues provides compelling evidence that the resulting candidate genes are worthy of further biomarker research.

In addition to addressing the central hypothesis, additional contributions are made to the field of colorectal neoplasia gene expression profiling. These contributions include:

The first systematic analysis of gene expression in non-diseased tissues along the colorectum To better understand the range of gene expression in non-diseased tissues, RNA extracts taken from along the longitudinal axis of the large intestine were studied.

The development of quality control methodologies for high dimensional gene expression data Complex data collection platforms such as oligonucleotide microarrays introduce the potential for unrecognized confounding variables. The exploration of quality control parameters across five hundred microarray experiments provided insights about quality control techniques.

The design of a custom microrray comprised of oligonucleotide probesets hybridising to RNA transcripts differentially expressed in neoplastic colorectal specimens A custom design oligonucleotide microarray was designed and tested combining the results of multiple biomarker discovery projects.

Introduction of a method to filter differentially expressed genes during discovery that may improve validation efficiencies of biomarker discovery based on gene expression measurements Differential expression discovery research is typically focused only on quantitative changes in transcript concentration between phenotype contrasts. This work introduces a method for generating hypotheses related to transcripts which may be qualitatively "switched-on" between phenotypes.

Identification of mRNA transcripts which are differentially expressed between colorectal adenomas and colorectal cancer tissues Transcripts differentially expressed between adenomatous and cancerous RNA extracts were discovered and then tested in independent tissues.

In conclusion, these results confirm the hypothesis that gene expression profiling can discriminate colorectal neoplasia (including adenomas) from non-neoplastic controls. These results also establish a foundation for an ongoing biomarker development program.

Declaration

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

.....

Lawrence Charles LaPointe

Acknowledgements

Firstly, I would like to thank my supervisors Prof. Graeme Young of Flinders University, Dr. Rob Dunne of CSIRO Mathematical and Information Sciences and Dr. Peter Molloy of CSIRO Molecular Health Technologies. I am indebted to Peter Molloy for reminding me that good science requires precision and careful consideration and that patience is often rewarded. I am grateful to Rob Dunne for teaching me skills that I will use for the rest of my career and for his excellent instruction of complex subject matter. I express my greatest thanks to Graeme Young, without whose guidance I would not have been able to start, conduct, or complete this research.

Collectively, my supervisors' guidance, scientific instruction, and ability to provide insightful criticism made this work possible.

I would like also to thank Clinical Genomics Pty Ltd and Enterix Inc for support of this research, including providing me ample time to dedicate to this study. In particular, I thank Howard Chandler, Max Mawhinney, and Peter Horrobin who have shared my vision that good science makes good business. With their support, I have been able to invest considerable time and energy into this research.

I thank my wife and family for love and support. I especially thank Karen for enduring my absence, inattention, and stress through these years without a single word of objection. Thank you for helping me to make this investment.

Finally, I express my deepest gratitude to the nameless patients and volunteers whose generous gift of clinical specimens forms the cornerstone of this research. To these individuals: your decision to contribute to the benefit of others even while you are confronted by the tragedy of colorectal cancer is inspirational. This thesis is aimed at discovering biomarkers which I hope will help others avoid your pain and I dedicate this work to you.

Chapter 1

Introduction

The science of cancer biomarker discovery for screening, diagnostic and therapeutic use has entered the "omic" era of biology [Weinstein, 2001]. Whether by "genomics", "proteomics", "epigenomics" or -omics to come, science by discovery-based techniques has become established as a legitimate alternative to traditional hypothesis-based methodologies [Ransohoff, 2003]. Fuelled by rapid developments in bioinformatics, the biologist's toolbox has become transformed from one-gene, one-protein experiments to gene expression profiling and data mining experiments, often analysing tens of thousands of genes [Liefers and Tollenaar, 2002]. The literature of colorectal cancer biomarker research has followed this transition over the last decade [Nannini et al., 2008].

Nevertheless, biomarker candidates from discovery-based research must be rigorously validated if they are to be clinically useful [Ransohoff, 2004a, Simon et al., 2003, Markowitz and Winawer, 1999]. This validation testing, ideally using clinically independent specimens, involves hypothesis testing. A decade of discovery-based science has demonstrated that failure to scientifically test hypothetical biomarker candidates often results in biomarkers of limited clinical value [Iaonnidis and Ntzani, 2003].

Given the ample scientific literature related to gene expression profiling in colorectal tissues (reviewed in Chapter 2), there is reason to believe that colorectal cancer tissues exhibit differentially expressed genes compared to non-neoplastic

tissues. Without effective validation testing, however, there is poor understanding about whether these differential gene expression patterns are reproducible in general for a wide sample of cancer specimens and, if so, whether these patterns will be specific to neoplastic tissues so as to discriminate non-neoplastic controls. Furthermore there is little knowledge about whether there are neoplastic gene expression patterns which are common to both benign colorectal adenomas and also malignant colorectal cancers.

This thesis explores the hypothesis that gene expression biomarkers can be used to accurately discriminate colorectal neoplastic tissues (both adenomas and cancers) from non-neoplastic controls.

If this hypothesis is shown to be correct, the key outcomes of this work will be to identify candidate biomarkers for colorectal neoplasia for future assay development and testing. To explore this hypothesis the first step is to understand whether genes are differentially expressed between neoplastic and non-neoplastic colorectal tissues in well designed scientific experiments involving careful validation testing of differential gene expression hypotheses. Next, if there is evidence of gene expression variation between phenotypes of interest, then what genes should be selected for further study including e.g. assay development, clinical testing, etc.?

The formal experimental elements of this research followed a two-phase strategic discovery and validation based approach:

1. High dimensional gene expression data were analysed to construct candidate biomarker hypotheses.
2. These biomarker hypotheses were then formally tested using an independently derived set of clinical specimens.

To test gene expression candidates mined from gene expression data, a custom-designed microarray was employed which included a large number of hypothetical markers. This project was designed to take a "multiplexed", hypothesis-based approach. Whereas traditional methods test putative candidates one at a time for diagnostic potential, this research analyzed thousands of potential markers in each observation.

1.0.1 Colorectal neoplasia

Colorectal cancer is unique among internal cancers in that early disease detection, through screening using simple tests that preselect patients who undergo diagnostic colonoscopic examination, has been shown to reduce cancer incidence [Mandel et al., 2000]. In effect, screening provides a powerful approach to prevention and cure.

Colorectal cancer is the only cancer for which there is “Level One” (randomised control trials) population evidence for reduced mortality and morbidity in persons of undefined risk by screening [Hardcastle et al., 1996, Kronborg et al., 1996, Mandel et al., 1993]. Furthermore, in a sixteen year follow up of the Minnesota Trial, Mandel demonstrated a reduction in the incidence of cancer in the cohort screened with faecal occult blood tests relative to the control population [Mandel et al., 2000]. This reduction in incidence is attributed to the detection and removal of precancerous polyps called adenomas in colonoscopic follow-up of screening positive patients [Winawer et al., 2006, Mandel et al., 2000].

Notwithstanding these findings, the current screening methods fall short of the diagnostic ideal, especially in the context of their limited ability to identify who is likely to have precancerous colorectal adenomas, and so triage people more efficiently to colonoscopy. Simpler tests that more accurately identify neoplastic lesions and which are more accurate than the guaiac based faecal occult blood tests (FOBTs) used in the randomised control trials are needed. Faecal immunochemical tests for human haemoglobin are an improvement over traditional guaiac tests, remain relatively inexpensive, and are more convenient,

but they are not sufficiently sensitive for adenomatous polyps and like guaiac FOBTs, detect all colorectal bleeding conditions, not just neoplasia [Levin et al., 2008]. While colonoscopy has excellent sensitivity for cancer and advanced adenomas, the procedure is costly, invasive and not without procedural health risk [Pickhardt et al., 2004, Lieberman, 2004, Hassan et al., 2008, Whitlock et al., 2008].

1.0.2 Adenomas as a target for cancer prevention

Diagnostic biomarkers are used to refine the risk profile of a given individual for significant disease [Markowitz and Winawer, 1999, Burt, 1996, Day, 1981]. Risk refinement is appropriate and useful as a means of selecting individuals to undergo costly, invasive and resource-limited treatments [Sachs, 2003]. In the field of colorectal cancer screening, one-step screening by colonoscopy or two-step screening using a simple test to select who gets colonoscopy is recommended [Levin et al., 2008]. Appropriate treatment may be undertaken at colonoscopy (polypectomy) with medical or surgical follow-up as otherwise necessary [Young et al., 1997].

As both a gold-standard diagnostic and therapeutic modality, colonoscopy provides a convenient endpoint: To select from a population (normal-risk or otherwise) those individuals that will most benefit from a diagnostic and possibly therapeutic colonoscopy [Lieberman, 2004]. Selection for colonoscopy provides both a theoretical framework for clinical utility and a practical guidepost for understanding the appropriate biomarker design inputs. Based on this criterion, the minimal acceptable positive threshold for a candidate biomarker should be the identification of such disease states that are detectable and, if possible, treatable by colonoscopy.

In the progression of colorectal cancer from the earliest mucosal changes to late stage metastatic disease, the adenoma represents the earliest stage of significant neoplasia worthy of clinical intervention by colonoscopy [Rex, 2002]. Identification and removal of colorectal adenomas not only lowers the morbidity and

mortality associated with colorectal cancer but also lowers the disease incidence as well [Winawer et al., 2006, Mandel et al., 2000]. By removing a precancerous, adenomatous tumour *before* the tumour exhibits the malignant adenocarcinoma phenotype, we provide the means to prevent cancer and lower the disease incidence. This opportunity has now been recognised by the US Multi-Society Task Force on Colorectal Cancer (including the American Cancer Society) who recently updated that body's screening guidelines to shift the emphasis away from diagnosis of curable cancer to prevention of the disease by highlighting the need to identify and remove tumours at the precancerous stage [Levin et al., 2008].

This research attempts to discover and select candidate biomarkers for colorectal neoplasia with sensitivity for both colorectal adenomas and colorectal cancer.

Finally, an optimistic but nonetheless promising view of technological progress suggests that in the future medical science will provide prophylactic or therapeutic treatments to those individuals who suffer from a predisposition to, or increased risk for, cancer. Such increased risk could also include individuals who possess a potentially dangerous precancerous lesion or predisposition without manifestation of any colonoscopy-detectable clinical symptoms. In other words, biomarkers might reflect the pre-neoplastic state and so define not only who has adenomas but who is most likely to develop adenomas. Hypothetically, were technology able to identify a pre-polyp "field effect" which will progress to malignancy with certainty, such a case might appropriately deserve some equally hypothetical treatment. Our acceptance of colonoscopy as the contemporary gold-standard diagnostic, however, necessarily limits any effective positive diagnosis to those cases where a colonoscopy can a) verify disease presence and b) ideally act as a positive intervention to neoplasia. Therefore, the aims of this thesis can be understood as being to identify candidate adenoma biomarkers that will be subject to the following (post-thesis validation) criterion: Candidate markers will ultimately be validated in typical screening populations for sensitivity and specificity for neoplasia (including precancerous adenomas and

cancer) that is detectable and, if possible, wholly treatable by colonoscopy.

Validation of candidate adenoma biomarkers will likely involve the design and development of *in vitro* diagnostic assays to measure either the mRNA biomarkers described in this work or a biologically related analyte such as, for example, a polypeptide translated from an mRNA transcript discovered here. Such analytes will be measured in a clinical specimen collected by either a doctor or the patient. The most convenient and non-invasive specimens relevant to colorectal neoplasia assays are faeces and blood.

Faecal sampling for colorectal cancer biomarkers is well established and has been routinely employed in faecal occult blood testing for more than 65 years [Schiff et al., 1942]. Current immunochemical assays used in colorectal cancer screening are based on detecting blood breakdown products released into the lumenal faecal flow as evidence of the dysplastic progression which occurs during carcinogenesis [Young et al., 1997]. While most cancers are believed to bleed as a consequence of this dysplasia, neoplasia is not the only clinical condition that causes colorectal bleeding bleeding. The possibility of non-neoplastic bleeding leads to a poor specificity associated with this assay technology [Smith et al., 2006].

Many of the biomarkers identified here are potentially expressed (as either RNA transcripts or otherwise) in colorectal tissues originating from colonocytes and surrounding stroma. Effective faecal sampling is thus predicated on the theory that neoplastic colorectal tissues will shed cells and/or molecules into the faecal flow. There is evidence that colonocytes and mucosal derived molecules can be found in the lumen and faeces [Nair et al., 2003, Lagerholm et al., 2005, Loktionov, 2007].

Alternatively, biomarkers discovered here could potentially be measured in circulation via transmission of molecules from neoplastic colorectal tissue into blood [Huang et al., 2003, Guadagni et al., 2001].

In either faecal or blood-based assays, complex specimen matrix effects including rapid degradation, protein masking, analyte dilution, etc. will introduce chal-

lenges to assay design. To improve the likelihood of successfully identifying and validating biomarkers for colorectal neoplasia, the gene expression data used here were measured using RNA extracts from freshly frozen colorectal tissue. Once the biomarker candidate "leads" have been identified with confidence at the tissue level, identifying an analyte of interest (e.g. nucleic acids or polypeptides) in a clinically useful specimen such as blood or faeces should become easier.

Chapter 2

Review of Colorectal Gene Expression

The introduction of high dimensional gene expression measurement systems such as oligonucleotide and cDNA microarrays has contributed significantly to the understanding of gene expression in a number of disease and non-disease systems [Ransohoff, 2004b].

The aim of this chapter is to review the literature related to gene expression patterns in the healthy adult colon as well as the major gene expression pathways associated with colorectal neoplasia. The literature of gene expression in the large intestine is broadly reviewed to establish a foundation for understanding gene expression in the colorectum. Next the biological literature related to adenoma formation is discussed with a focus on the natural history of colorectal oncogenesis relevant to identification of molecular biomarker diagnostics. An overview of the rapidly growing body of gene expression data collected in colorectal tissues is then presented. Finally, the central hypothesis of this thesis, that gene expression biomarkers can be used to accurately discriminate colorectal neoplastic tissues from non-neoplastic controls, is framed in the context of this literature review.

2.1 Gene expression in the large intestine

Gene expression patterns in the colorectal mucosa reflect underlying programming that is defined and maintained through balanced forces at the level of tissue, cells and molecules. The adult colorectal epithelium undergoes constant turnover as terminally differentiated, non-dividing cells are shed, or perhaps absorbed, at the luminal epithelial surface and are replaced by a wellspring of new cells originating from the crypt base [de Santa Barbara et al., 2003, Gordon and Hermiston, 1994, Loktionov, 2007, Booth and Potten, 2000]. This process of constant regeneration is achieved by mucosal epithelial cells exhibiting a prototypical pattern of proliferation, differentiation, migration and apoptosis along the crypt axis [Mariadason et al., 2002]. In the healthy colorectum these cellular dynamics are modulated by a mix of endogenous cell cycle and adhesion molecules as well as by complex exogenous signalling between the epithelium and underlying mesenchymal stroma [Plateroti et al., 1998]. Many of these signals maintain the healthy colorectal mucosa phenotype by regulating transcriptional control of gene expression cascades [Radtke and Clevers, 2005].

2.1.1 Gene expression patterned during organogenesis

In many vertebrates, including humans, the fully formed adult colorectal epithelium phenotype develops after birth, perhaps in response to weaning or other extrinsic signal sources [Duluc et al., 1993]. This process includes the disappearance (or flattening) of small-intestinal-like villi as well as the transient expression of small intestinal enzymes such as brush-border hydrolases [Foltzer-Jourdainne et al., 1989]. Starting after birth and continuing throughout life, the crypt-lumen axis of the epithelium undergoes continuous regeneration. There is strong evidence that this constant epithelial renewal is dependent on mesoderm-derived signals although there may also be autonomous epithelial-specific development in some regions of the gut [Kedinger et al., 1998]. For example, the midgut endoderm may, in part, be self-determined by endogenous signals [Duluc et al., 1994]. Further, there is also evidence that signalling may be bi-directional with

the endoderm capable of inducing differentiation of nonsplanchnic mesoderm e.g. to develop smooth muscle [Roberts, 1999].

The embryological gut is lined by a single layer of cuboidal-columnar endoderm epithelium surrounded by a sheath of splanchnic mesoderm. As the mesoderm differentiates into smooth muscle along the gut, the primitive gut tube develops into clearly delineated regions along the anterior-posterior axis to form the fore-, mid-, and hindgut [Roberts, 1999]. Intestinal organogenesis involves the formation of a monolayer epithelium in the basal crypts [Duluc et al., 1994] where endogenous and exogenous signals stimulate proliferation of stem cells and differentiation along two colorectal epithelial cell lineage pathways [Radtke and Clevers, 2005]. Most cells differentiate to become absorptive enterocytes while the rest develop secretory functions [Marshman et al., 2002]. The secretory lineage includes both the mucus secreting goblet cells and hormone producing enteroendocrine cells [Peifer, 2002, Radtke and Clevers, 2005].

2.1.2 Expression along the proximal-distal axis

The traditional division of the human colorectum into proximal and distal regions divided approximately two-thirds along the transverse colon is supported by the embryology of the large intestine. While the proximal large intestine develops from the embryonic midgut and is perfused by the superior mesenteric artery, the distal large intestine forms from the embryonic hindgut and is supplied blood from the inferior mesenteric artery [Babyatsky and Podolsky, 2003, Iacopetta, 2002]. In a clinical context, this division is useful because of the way in which clinical diseases are differentially exhibited along the length of the colorectum.

Proximal-distal patterning of the primitive gut tube is partly controlled by homeobox genes [James and Kazenwadel, 1991, Montgomery et al., 1999, Booth and Potten, 2000]. The four groups of *HOX* gene paralogues (39 members) consist of highly conserved transcription factors that specify the identity of body segments along the anterior-posterior axis of the developing embryo and are known body

plan regulators in many organisms including Drosophila and humans [Hostikka and Capecchi, 1998, Kosaki et al., 2002, Montgomery et al., 1999]. In the gut endoderm, sonic hedgehog (*SHH*) has been shown to be an upstream activator of both *HOXD* and *BMP4* in mesoderm where ectopic over-expression of *SHH* can result in an over-proliferation of the gut-specific mesoderm. The transcription factor forkhead (*Fkh*), which is required for fore- and hindgut development in *Drosophila*, may, in turn, be an upstream activator of *SHH* expression leading to the regionalisation of the sub-adjacent mesoderm after elongation [Roberts, 1999].

Studies suggests that there is a distinction between the gene expression patterns of proximal colonic tissues and distal colorectal tissues [Glebov et al., 2003, Komuro et al., 2005, Birkenkamp-Demtroder et al., 2005]. To explore patterns of gene expression along the proximal-distal axis of the large intestine, we used full-genome microarrays to build expression profile "maps" that identify individual genes whose expression appears to be location dependent as well as to characterise the nature of gene expression change along the proximal-distal axis.

Work discussed in Chapter 6, shows that transcript abundance follows two broad patterns along the proximal-distal axis of the large intestine. The dominant pattern is a proximal-distal expression pattern consistent with the midgut-hindgut embryonic origins of the proximal and distal gut, with a sharp transition between the ascending and descending colon. A secondary pattern is characterised by a gradual change in transcript levels from the cecum to the rectum, nearly all of which exhibit increasing expression toward the distal tissues.

2.2 Gene expression along the crypt-axis

The control of tissue organisation along the crypt axis involving proliferation, differentiation, migration, senescence, anoikis and apoptosis is fundamental to both the continual regeneration of the healthy colorectal mucosa and to understanding the propensity for neoplastic growth that stems from a lack of balance of molec-

ular control [Liotta and Kohn, 2001]. Four primary gene expression pathways have been shown to play a role in creating and maintaining healthy crypt-lumen axis dynamics: the Wnt pathway regulating Tcf-4/Lef-1 downstream elements, the TGF-β/BMP pathway, Notch signalling, and the Hedgehog pathway [Klaus and Birchmeier, 2008]. These pathways have also been shown to interact co-operatively to maintain the colorectal phenotype. For example Wnt and Notch appear to cooperate in control of stem cell self-renewal [He et al., 2004]. These pathways are reviewed here.

2.2.1 Wnt signalling

Wnt proteins are secreted growth factors regulating cell fate during development that are conserved across multicellular animals [Miller, 2002]. In addition to playing a key role in organogenesis during embryological development, Wnt signals issued by the mesenchymal cells below the crypts of intestinal villi have been shown to play a role orchestrating the carefully balanced epithelial regeneration process [Peifer, 2002, van de Wetering et al., 2002, Batlle et al., 2002]. Downstream signal transduction of this pathway acts through β-catenin and the Tcf/LEF transcription factors to stimulate gene expression of target genes of this pathway. Importantly, mutations or disruptions that lead to over-expressed Wnt targets may be sufficient to induce adenoma development, but aberrant Wnt signalling alone does not lead to carcinoma (reviewed in Ilyas et al. [1999b] and Narayanan et al. [2003]).

The cascade of genes that are directly or indirectly expressed or repressed by downstream Wnt-dependent Tcf/LEF transcription activation appear to provide the potential for neoplastic growth. The list of target genes includes key cell-cycle regulator genes, differentiation controls, morphological and adhesion molecules, angiogenesis stimulators, and suppressors of apoptosis. In the small intestine, the effects of the Wnt pathway appear to be balanced by the crypt boundary created by Eph-ephrin interactions [Batlle et al., 2002]. If, as expected, a similar mechanism is involved in the colorectum, then the Wnt pathway may

be actively involved in maintaining proliferation and the crypt-axis morphology in the healthy colon.

Canonical Wnt pathway

In the absence of other factors, the cell-cell adhesion molecule cytoplasmic β-catenin (*CTNNB1*) is competitively bound to either E-cadherin at the cytoplasmic surface or to APC which selectively targets β-catenin for ubiquitin degradation [Aberle et al., 1997, Mann et al., 1999, Pierce et al., 2003]. Consequently, β-catenin levels in a normal adult cell are generally quite low [Munemitsu et al., 1995]. APC plays a critical role in the degradation by anchoring and stabilising β-catenin as part of a large complex that includes Axin, GSK3β, Dishevelled (Dsh), and GSK-binding protein (GBP-Frat). β-catenin anchored to this complex is then N-terminally phosphorylated by GSK3β and casein kinase 1α at four conserved Ser/Thr phosphorylation sites which results in subsequent binding of β-catenin to the F-box protein β-transducin repeat-containing protein (β-TrCP) [Barker et al., 2001, Aberle et al., 1997]. β-TrCP, a member of the ubiquitin ligase complex, ubiquinates and targets β-catenin for degradation by proteasomes [Roose and Clevers, 1999, Miller, 2002, Ougolkov et al., 2004, Klaus and Birchmeier, 2008].

Wnt signalling (e.g. during embryogenesis) is initiated by Wnt protein binding to cell surface co-receptors Frizzled (Fzd) and LRP. This leads to a phosphorylation of Dsh which then antagonises GSK3β, preventing the proximity phosphorylation and subsequent ubiquitin degradation of β-catenin. As the cytosolic concentration of β-catenin rises, β-catenin begins to complex with the Tcf/LEF family of HMG-domain transcription factors and is then shuttled to the nucleus [Korinek et al., 1997, Miller, 2002, Mann et al., 1999]. The nuclear β-catenin-Tcf/LEF complex may initiate target transcription by recruiting Brg-1 (Brahma-related gene1, an ATP-dependent component of the SWI/SNF and Rsc chromatin remodelling complex) to remodel chromatin near the Tcf target genes [Barker et al., 2001]. Without such remodelling, “closed” chromatin

is inaccessible to transcription factors and unable to bind basal transcription machinery such as RNA pol II [Laybourn and Kadonaga, 1992].

Target genes of the Tcf/LEF family of transcription factors were determined by gene disruption studies, inducible promoter models, etc. that provide strong evidence for transcriptional initiation following DNA binding of the Tcf/LEF complex. These are shown in Table 2.1.

Table 2.1: Genes that demonstrate up-regulation by Tcf/LEF

c-Myc	Nuclear DNA binding protein	[He et al., 1998]
CCDN1	Cyclin D1	[Tetsu and McCormick, 1999] [Arber et al., 1996] [Shtutman et al., 1999]
MMP7	matrilysin (matrix metalloproteinase)	[Brabletz et al., 1999] [Crawford et al., 1999]
fra-1	Transcription factor	[Mann et al., 1999]
c-jun	Transcription factor	[Mann et al., 1999]
uPAR	urokinase-type plasminogen activator	[Mann et al., 1999]
LEF1&TCF7	Transcription activator proteins	[Roose and Clevers, 1999] [Roose et al., 1999] [Hovanes et al., 2001]
axin2	Axin, cytoskeletal components, negative regulatory components of the β-catenin induced pathway.	[Yan et al., 2001]
hNkd	human homologue of mouse Nkd, with axin2 acts as negative regulator of Tcf/LEF pathway.	[Yan et al., 2001]
BMP4	Bone morphogenetic protein 4 – member of the TGF-β superfamily of growth factors perhaps involved with differentiation.	[Kim et al., 2002]
ITF-2	immunoglobulin transcription factor 2	[Kolligs et al., 2002]
PPARδ	Peroxisome proliferator-activated receptor delta (nuclear receptor that acts as a ligand dependent transcription activator)	[He et al., 1999]

Continued on Next Page...

Table 2.1 – Continued

NBL4	Novel band 4.1-like protein 4, member of a family of proteins that could have a role in metastasis. Thought to be involved in regulating interaction of the cell cytoskeleton and plasma membrane.	[Ishiguro et al., 2000]
Nr-CAM	neuronal cell adhesion molecule; a transmembrane cell adhesion protein mostly expressed in normal cells of the nervous system.	[Conacci-Sorrell et al., 2002]
VEGF	vascular endothelial growth factor	[Zhang et al., 2001]
CD44	a family of cell-surface glycoproteins generated from a single gene by alternative splice variants and glycosylation; could promote cell motility and growth.	[Wielenga et al., 1999]
Survivin	Suppresses apoptosis.	[Zhang et al., 2001] [Kim et al., 2003b]
ENC1	Ectodermal Neural Cortex 1, may lead to neoplasms by preventing regulated differentiation of colonic mucosae and neural cells (related to morphology control)	[Fujita et al., 2001]
CLDN1	claudin1	[Miwa et al., 2000]
gastrin		[Koh et al., 2000]
Genes that demonstrate down-regulation by Tcf/LEF		
ZO-1	zona occludens	[Mann et al., 1999]
MCP-3	Monocyte chemotactic protein 3, may disturb differentiation in colon cells	[Fujita et al., 2000]
DRCTNNB1A	Down regulated by CTNNB1A	[Kawasoe et al., 2000]
KIAA1199	Novel tumour marker	[Sabates-Bellver et al., 2007]

MYC (alias *c-Myc*) was one of the first genes to be shown to be activated through Tcf/LEF signalling following earlier studies that showed this gene to be up-regulated in colon cancers [Finley et al., 1989, Rochlitz et al., 1996, He et al., 1998]. This oncogene is a member of the MYC family and encodes a small nuclear DNA-binding protein regulating proliferation, transformation, and differ-

entiation [Aiello et al., 2004]. As a promoter of neoplastic growth, among many other functions, *MYC* has been shown to directly repress the p21CIP1/WAF1 promoter which mediates cell cycle G1 arrest and differentiation through cyclin-dependent kinase inhibition [van de Wetering et al., 2002].

The Tcf/LEF transcriptional activation complex may also directly stimulate neoplastic transformation through activation of cyclin D1 *CCND1* [Tetsu and McCormick, 1999, Shtutman et al., 1999]. Accumulating cyclin D1 associates with the cyclin-dependent kinases to create a catalytic complex that phosphorylates the retinoblastoma tumour suppressor protein (Rb), freeing E2F to initiate cell G1 phase progression [Turner et al., 2000]. Consequently, cyclin D1 occupies a crucial role in cell cycle regulation and its constitutive activation by Wnt signalling could conceivably shift the cell out of proliferative balance [Smalley and Dale, 1999, Waltzer and Bienz, 1999].

Other Tcf/LEF transcription targets that could have a disruptive effect on the cell cycle balance, include *JUN* (c-jun),*FOSL1* (fra-1), and *PPARD* (PPAR δ). The genes for bone morphogenetic protein 4 (*BMP4*) (discussed further below) and ectodermal neuronal cortex 1 (*ENC1*) are also transcriptionally activated downstream from Wnt and are likely involved in regulating cell differentiation [Kim et al., 2002, Nishanian et al., 2004, Fujita et al., 2001].

The activation of *BIRC5* (survivin), an inhibitor of apoptosis, by Tcf/LEF further blurs the line between normal mucosa and adenomatous polyps. There is reasonably strong evidence that survivin is expressed in the lower crypts of normal mucosa with decreasing expression toward the luminal surface [Zhang et al., 2001]. This expression pattern is inversely correlated with wild-type APC expression along the crypt axis, leading Zhang et al. [2001] to postulate that wtAPC-induced suppression of survivin may cause stem cell progeny to lose their apoptosis-resistant phenotype as they migrate upwards [Zhang et al., 2001, Kawasaki et al., 2001, Lin et al., 2003]. Conversely, aberrant constitutive Wnt signalling through Tcf/LEF may provide an intrinsic mechanism for these migrating epithelial cells to avoid the natural death pathway.

In addition to these intrinsic effectors of neoplastic potential, several of the Tcf/LEF gene targets could mediate extrinsic factors involved with neoplasia. For example, *MMP7* (matrilysin) and urokinase-type plasminogen activated receptor (*PLAUR*) may potentiate extracellular invasion by matrix proteolysis [Mann et al., 1999, Brabletz et al., 1999, He et al., 1999]. In a review of matrix metalloproteinases, Chambers and Matrisian suggest an expanded role for proteinases such MMP7 through regulation of the growth environment around the primary tumour by providing access to growth factors from the extracellular matrix and by assisting angiogenesis [Chambers and Matrisian, 1997].

Given the central role Wnt signalling plays in tissue development and maintenance, it is not surprising that there may be self-regulating feedback mechanisms in response to Tcf/LEF activation. *AXIN2* and *NKD1* are both activated by Tcf/LEF and are also presumed to be negative regulators of Wnt signalling [Yan et al., 2001]. On the other hand, Hovanes et al. [2001] have shown that β-catenin-Tcf complexes also selectively activate LEF-1 isoforms and avoid a second, dominant-negative, form thereby inducing a positive feedback loop when unregulated high concentrations of β-catenin accumulate in cancer. This work by Hovanes et al. extends and builds on earlier work by Roose et al. [1999] showing that Tcf1 (encoded by *TCF7*) was similarly feed-forward activated.

Finally, at least two genes appear to be induced by the Wnt pathway in a Tcf/LEF-independent manner. *WISP1* (Wnt-1 induced secreted protein) is a member of a family of growth factors that mediate the growth signals between the epithelial tumour cell and the surrounding stromal cells [Pennica et al., 1998]. Xu et al. [2000] showed that *WISP1* was not stimulated by β-catenin nuclear accumulation but appears instead to be up-regulated by some Tcf/LEF-independent manner. Xu et al. postulate that *WISP1* may be induced by the intermediary affects of β-catenin on cyclic AMP leading to activation of protein kinase A and phosphorylation of CREB protein with downstream transcription. This complex relationship underscores the inter-dependent and complex nature of potential molecular networks involved with colorectal transformation.

MLLT6 (also known as AF17), a fusion partner of the MLL gene in acute

leukaemia, has also been shown to be activated downstream of the Tcf/LEF complex without direct interaction between the *MLLT6* transcription activation site and Tcf/LEF factors [Lin et al., 2001]. Experimental evidence suggests that MLLT6 is involved in cell-cycle proliferation similar to c-Myc and cyclin D1.

Non-canonical Wnt

In addition to the canonical Wnt-β-catenin signalling pathway described above there is also a second pathway to activate Wnt-transcriptional targets. The non-canonical Wnt pathway acts through the usual Frizzled and Dishevelled receptors in a β-catenin independent manner. The planar cell polarity and Ca^{2+} pathways are examples of non-canonical Wnt signalling. However it is important to note that mutations resulting in constitutive non-canonical Wnt signalling in human cancers have not been observed [Klaus and Birchmeier, 2008].

2.2.2 TGF-β Superfamily

The TGFβ superfamily is involved in regulation of a broad continuum of cell processes including proliferation, differentiation apoptosis, matrix modelling, and migration. The superfamily includes secreted cytokines such as the TGFβ isoforms, bone morphogenetic protein (BMP), Nodal, growth and differentiation factors, and activins [Radtke and Clevers, 2005, Beck et al., 2006, Ilyas et al., 1999a, Ross and Hill, 2007]. While the central signal cascade follows a relatively straightforward signal transduction chain, the final step of DNA-binding to effect transcriptional changes (positive and negative) involves recruitment of cell-type specific and cell-context-specific co-factors that provides means for a rich, complex physiology for this pathway [Ilyas et al., 1999c, Ross and Hill, 2007].

In the healthy adult colon, the TGFβ pathway appears to counteract the proliferative signals of the Wnt pathway, with epithelial growth suppression re-

sulting from ligand and receptor activity in the differentiated compartments of the crypt-villous axis [Radtke and Clevers, 2005]. By blocking cell cycle progression through various biochemical and molecular interventions (described below), TGFβ signalling stabilises the epithelial phenotype of the higher crypt terminus. In this respect TGFβ acts in tumour suppressing role [Matsuzaki and Okazaki, 2006].

Mechanisms of TGFβ superfamily signalling

Signal transduction of the TGFβ cascade is initiated when a TGFβ ligand binds to a receptor complex of Sarine-threonine kinase type I and type II receptors. The ligand binds via the type II TGFβ receptor which then phosphorylates the type I TGFβ-receptor which then transduces the signal into the cytoplasm by in turn phosphorylating one of the signal transduction Smad proteins (Smad-1, -2, -3, -5, or -8). In the phosphorylated state this receptor Smad, or R-Smad, either homodimerises with a second R-Smad or forms a heteromeric dimer with the common Smad (Smad-4). Finally, this activated (homo- or hetero-) dimer translocates from the cytoplasm to the nucleus, where the complex regulates transcription of target genes by either activation or repression [Hahn et al., 1996, Moskaluk and Kern, 1996, Riggins et al., 1996, Ross and Hill, 2007].

This relatively simple signal cascade becomes more complicated in the nucleus as the R-Smad dimer interacts with a range of transcriptional activators and repressors to effect a wide ranging gene expression program [Ross and Hill, 2007]. For example, downstream, the transcription factors repress cell cycle progression beyond G1 by inducing expression of *CDKN2B* (alias: P15,P15INK4B) and *CDKN1* (p21,P21CIP/WAP) and by down- regulating *CDC25A* cdc25A. The regulation of these targets results in downstream inhibition of various cyclin dependent kinases (cdk4, cdk6, and cdk2) to block Rb phosphorylation(reviewed in Massague et al. [2000] and Radtke and Clevers [2005]). TGFβ inactivation of G1 Cdks, however, can likewise be blocked by c-Myc which has been shown to interfere with the rapid activation of *CDKN2B* and *CDKN1*. To overcome this inhibition, TGFβ acts through a secondary pathway to directly down-regulate

c-Myc transcription [Warner et al., 1999, Pietenpol et al., 1990, Malliri et al., 1996].

Many of the co-factors that interact with the R-Smads are themselves under regulatory control, which allows the cell to restrict the context of TGFβ signalling. The TGFβ pathway can also directly inhibit itself through binding of one of the *inhibitory* Smads, i-Smad-7 or iSmad-8 [Ross and Hill, 2007]. Finally, there is also evidence that TGFβ may stimulate anti- proliferative (growth control) signals by a Smad-4 independent pathway involving the MAP kinases JNK, p38, and Erk [Massague et al., 2000].

2.2.3 Notch control of lineage differentiation

The Notch signalling pathway is highly conserved and plays a role directing the cell differentiation program mediated by cell-cell contact [Koch and Radtke, 2007]. There are two signalling cascades which follow from Notch signalling. The canonical notch signalling perpetuates the stem cell phenotype and, along with Wnt signalling, is required to maintain the crypt compartment [Katoh and Katoh, 2007, Koch and Radtke, 2007]. The non-canonical pathway, on the other hand, stimulates differentiation and transcriptional activation [Katoh and Katoh, 2007].

Notch signalling involves four transmembrane receptors: NOTCH1, NOTCH2, NOTCH3, and NOTCH4 which can bind to the extracellular DSL domain from one of five transmembrane ligands: Delta-like 1 (DLL1), DLL3, DLL4, Jagged1 (JAG1), and JAG2. In addition to these ligands, there are three atypical ligands – so defined because they lack the usual N-terminal DSL domain – DNER, F3/Connectin and NB-3 [Katoh and Katoh, 2007].

In the bound state Notch receptors are proteolyticly cleaved to release Notch intracellular domain (NICD) by γ-secretase and metalloproteinase of the ADAM family [Radtke and Clevers, 2005, Koch and Radtke, 2007]. After release, NICD translocates to the nucleus and binds one of two transcription factors. In the

canonical pathway, NICD binds the transcription factor CBF1 and Mastermind complex, which results in the transcription of target genes such as *HES1*, a member of the HES family of transcriptional repressors. Thus the canonical pathway maintains the stem cell progenitor by repressing downstream transcription and differentiation via e.g. *HES1*. In the non-canonical pathway, NICD binds NF-κB which results in activation of NF-κB targets and differentiation away from the stem cell state [Katoh and Katoh, 2007].

In the colorectal crypt Notch also mediates *HES1* to control cell differentiation toward either absorptive enterocytes or secretory goblet cells. Following the canonical signalling pathway, HES1 expression represses *MATH1* resulting in absorptive enterocyte development. Conversely non-canonical Notch signalling bypasses HES1 resulting in *MATH1* expression and secretory goblet cell formation [Radtke and Clevers, 2005].

In addition to HES-family genes, other targets of Notch possibly include HERP transcription family, cyclin-dependent kinase inhibitor 1 (*CDKN1A*), cyclin-D1 (*CCND1*), Notch regulated ankyrin repeat protein (*NRARP*), deltex 1 homologue (*DTX1*), pre T-cell antigen receptor alpha (*PTCRA*), and the ubiquitin ligase *SKIP2* [Koch and Radtke, 2007].

2.2.4 Hedgehog Pathway

Unlike the first three pathways discussed here, the hedgehog pathway acts in a paracrine fashion, with signal peptides secreted from the epithelial cells binding and transducing expression in the mesenchymal sub-epithelial myofibroblasts and smooth muscle cells [Madison et al., 2005]. The hedgehog pathway plays a key role in growth and maintenance of crypt-villous architecture [Madison et al., 2005, Taipale and Beachy, 2001]. In the small intestine, inhibition of hedgehog signalling results in disrupted villus formation.

There are three mammalian hedgehog proteins, sonic hedgehog (*SHH*), Indian hedgehog (*IHH*), and desert hedgehog (*dhh*) [Madison et al., 2005]. Signal

transduction via these genes is based on repressive interactions [Taipale and Beachy, 2001]. Following intra-molecular cleavage and C-terminal ester bonding to cholesterol, hedgehog signals (Hh) are secreted for potential binding to the Patched family of receptors, PTCH1 and PTCH2 [Taipale and Beachy, 2001]. Binding to PTCH1 releases Smoothened transducer (SMO) which then in turn inhibits assembly of the GLI degradation complex resulting in GLI stabilisation and transcriptional activation of GLI targets [Katoh and Katoh, 2006]. A lack of hedgehog stimulation results in SMO inhibition by PTCH1/PTCH2, formation of GLI degradation complex and repression of GLI induced transcription.

GLI transcription targets include *GLI1*, *PTCH1*, *CCND2*, *FOXL1*, *CCND1*, *BMP*, *Wnt*, and *JAG2* [Katoh and Katoh, 2006, Bian et al., 2007].

While aberrant hedgehog signalling has been shown to be involved in a number of human cancers including basal cell carcinoma, medulloblastoma, and small cell lung carcinoma, prostate cancer, and pancreatic adenocarcinoma, hedgehog activation (or deactivation) in colorectal cancers is controversial [Chatel et al., 2007, Bian et al., 2007].

2.3 Molecular biology of Adenoma Formation

2.3.1 Cell cycle balance and oncogenesis

Colorectal cancer, like all cancers, is presumed to be the phenotypic reflection of genetic defects, i.e. genomics instability, leading to an out-of-balance state between the basic cell mechanisms of proliferation, DNA repair, differentiation and apoptotic growth regulation [Aiello et al., 2004, Hao et al., 1998]. In approximately 5% of cancers these defects have been shown to stem from inherited susceptibility observed in familial cancer syndromes [Ilyas et al., 1999b]. The majority of colorectal cancers are believed to be independent of a dominant genetic background and are thus called sporadic colorectal cancer.

The intrinsic elements of equilibrium in a “normal” cell – and disequilibrium in

cancer – include positive forces of proliferation, negative forces of cell cycle regulation and apoptosis, and forces that act in both positive and negative fashion. These elements all result from progressive genomic instability during oncogenesis. A table of these elements of control is shown below in Table 2.2. In addition to intrinsic elements, a neoplastic cell may also affect changes in the neighbouring tissue through extrinsic forces that can further propel a cell or tissue out of balance [Ilyas and Tomlinson, 1996, Augenlicht et al., 2002]. These elements are also shown in Table 2.2.

Table 2.2: Elements of Cell Cycle Balance

INTRINSIC ELEMENTS
Cell division membrane (self) presentation, e.g. effecting cell adhesion, migration differentiation status apoptosis and natural cell death molecular (DNA) repair (genomic and non-genomic) altered signalling
EXTRINSIC ELEMENTS
vascular and nutrient supply to tissue control of tissue structure (e.g. connective tissue.) intercellular signalling and growth factor response extracellular milieu (e.g. faecal stream, microflora)

A complementary view of cancer-related gene classification has been proposed by Kinzler and Vogelstein that identifies genes as either “gatekeepers” or “caretakers” [Kinzler and Vogelstein, 1997]. In this context, a gatekeeper gene refers to a gene that directly or indirectly is involved in cell proliferation, growth, restriction, or death (e.g. p53, Apc, Rb). Caretaker genes, on the other hand, function to maintain genetic integrity (e.g. mismatch repair (MMR) genes) and their mutation is likely to increase susceptibility to further mutational events. The interrelationship is crystallised by caretaker mutations that lead to gatekeeper mutations creating the potential for catastrophic results to the cell and tissue.

If a cancerous tissue is characterised by disequilibrium in the homeostatic processes that contain and limit cell proliferation, a cancerous cell is likewise sub-

ject to molecular events that trigger or increase this disequilibrium. Such events may arise from intrinsic or extrinsic forces. Intrinsic forces include replication errors which can become oncogenic if such errors escape DNA repair processes or are not removed by apoptotic deletion (i.e. cell death). Extrinsic sources of cell disequilibrium are induced by external mutagenic agents. These initiating trigger events (mutations or otherwise) lead to either gene disruption and loss-of-function or increased expression/activation resulting in a gain-of-function. Bronchud et al. [2004] list a number of molecular mechanisms that result in gene and/or protein product disruption and loss-of-function. These mechanisms are shown in Table 2.3

Table 2.3:

Mechanisms of Gene/Protein Disruption
Entire gene deletion
Loss of chromosome
Partial gene deletion
Disruption of gene structure (translocation/inversion)
Sequence insertion into the gene
Promoter mutation reducing mRNA levels
Decrease in mRNA stability
Inactivation of donor splice sites - exon skipped
Activation of cryptic splice sites
Frameshift translocation
Conversion of a codon to a stop codon
Replacement of an essential amino acid
Prevention of post-translational processing
Prevention of correct cellular localisation of product
Altered methylation of promoter

Unless the affected gene is density or concentration dependent, gene disruption leading to a loss-of-function is a recessive trait that remains phenotypically hidden because the second, wild-type, allele is adequate to protect the cell against oncogenesis [Ephrussi et al., 1969]. If, however, a second mutational event disrupts the wild-type allele (causing a loss-of-heterozygosity) of the gene in question, the combined mutational events may result in a loss-of-function. This phenomenon was first proposed in 1973 by Alfred Knudson while studying age-specific cancer incidence and is consequently referred to as “Knudson’s Two-Hit

Hypothesis" [Knudson, 1993, 1973, 1971]. The APC gene is a typical example of the two-hit hypothesis: mutation in one allele is not sufficient to lead to adenomatous polyp formation but loss of heterozygosity (LOH) through deletion of the second allele may initiate neoplasia [Klaus and Birchmeier, 2008].

Alternatively, the molecular event may result in a gain of function for a given gene or protein [Bronchud et al., 2004]. In principle, if one presumes that loss-of-function mutations will generally be observed as phenotypically recessive, then conversely, gain-of-function mutations will generally be phenotypically dominant. A hypothetical list of such mechanisms is provided in Table 2.4 and logically mirrors many of the "disruptive" mechanisms discussed above.

Table 2.4:

Mechanisms of Gain-Of-Function
Gene copy number increase
Chromosomal duplication
Translocation of promoter
Promoter mutation increasing mRNA levels
Increase in mRNA stability
Amino acid change conferring increased functionality
Error in post-translational processing
Incorrect cellular localisation resulting in increased activity
De-Methylation of promoter
Loss of imprinting

While the organogenesis and histology of the colorectum are beyond the scope of this review, one should bear in mind that the colorectal mucosa is under continuous pressure to regenerate as the epithelial surface of the gut lumen is sloughed away [Augenlicht et al., 2002]. This pressure naturally creates within the colorectal mucosa an equilibrium state that is relatively static at the tissue level, but dynamic at the cellular level. In other words, at the tissue level the total number of cells dividing within the crypt should equal the number of cells dying and being shed at the luminal surface (in the adult, non-growing, phenotype). Too few mitotic cells in the crypt would create an atrophied mucosal state and too many could initiate cellular hyperplasia. At the level of an individual cell however, there is shift of equilibrium along the crypt axis from proliferation in the lower crypt to terminal differentiation moving toward the mucosal surface.

To illustrate this dichotomy, note that nearly continuous, undifferentiated growth is a general hallmark of neoplasia including e.g. colorectal adenomas. On the other hand, this vague description also precisely includes the (regulated) division of stem cells found within the colonic crypts, without which the colon would cease to regenerate and die. Consequently, the efforts of this thesis to identify sensitive and specific biomarkers for neoplasia may be affected by the nature of transformation from normal to neoplastic and eventually to malignant (i.e. invasive) disease. In some cases, biological disruption by one or more of the mechanisms found in Tables 2.4 and 2.3 provides a clear delineation between healthy and disease states with a definable, if not discrete state change. Alternatively, this work may uncover diagnostic patterns of gene expression that reflect subtle, coincident perturbations of several otherwise normal regulatory pathways.

In his review of colorectal oncogenesis, Potter describes the analogous microscopic structure of the colorectum covered with microscopic crypts compared to the villous nature of the small intestine [Potter, 1999]. While the intestinal villi are presumed to increase the available surface area for nutrient absorption, the colonic crypts are unlikely to function in this manner. As an alternative, Potter suggests that this histological structure may provide the highly mitotic, undifferentiated stem cells with spatial separation from the mutagenic faecal stream passing through the lumen.

One can synthesise a conceptual framework based on these observations whereby subtle changes in the orchestration of regulatory mechanisms (e.g. proliferation, adhesion, or morphology) could potentiate further disruptive events by simply exposing the pluripotent crypt cells to the lumen contents. From a diagnostic perspective, this carcinogensis mechanism suggests the possibility that neoplasia could arise from subtle perturbations of the cellular control networks which have significant downstream effects. Nevertheless, the aim of this thesis will be to design appropriately strong validation protocols that will satisfy the minimal sensitivity requirements of colonoscopic detection, as discussed in Chapter 1.

2.3.2 The adenoma-carcinoma sequence

The progression of colorectal carcinoma through well defined and histologically distinct phenotypic stages provides a useful foothold for genetic study from initiation of oncogenesis (focal microscopic dysplastic lesions) through to formation of macroscopic adenomas and the eventual acquisition of the invasive phenotype, the hallmark of cancer (adenocarcinoma) first manifested as *in situ* carcinoma [Morson, 1974, Hill et al., 1978, Reale and Fearon, 1997]. Consequently, colorectal cancer has served as a general model of molecular oncogenesis beginning with the work of Muto et al. [1975] and continuing with the contribution of Vogelstein et al. [1988] [Fearon and Vogelstein, 1990]. The cascade of mutations associated with familial adenomatous polyposis (FAP) and microsatellite instability seen in hereditary nonpolyposis colorectal cancer (HNPCC) have provided strong theoretical and empirical evidence in relation to the disease mechanisms for the two canonical pathways of colorectal oncogenesis [Burt and Samowitz, 1988]. These canonical pathways are illustrated below in the prototypical "Vogelgram" depicting the adenoma-carcinoma sequence shown in Figure 2.1 [Soreide et al., 2008].

2.3.3 Disruptive Wnt signalling and neoplasia

Approximately 90% of sporadic colorectal neoplasias and 100% of FAP neoplasias exhibit aberrant, constitutive Wnt signalling [Giles et al., 2003]. In colon cancer, disruption of the Wnt pathway occurs frequently by mutations in the adenomatous polyposis coli gene *APC*. The identification of APC-associated mutations in FAP was one of the first cancer syndromes to be elucidated [Vogelstein et al., 1988]. Over 120 mutational hot spots have been identified in *APC* [Su et al., 1993], nearly all (95%) of which transcribe (nonsense) premature stop codons leading to translation of a truncated protein form [Nishisho et al., 1991]. Such truncations lack phosphorylation sites for GSK3β and/or binding sites for axin and β-catenin. *APC* can also be silenced by hyper-methylation [Hiltunen et al., 1997]. Any failure of APC to bind β-catenin prevents the ubiquitinisation

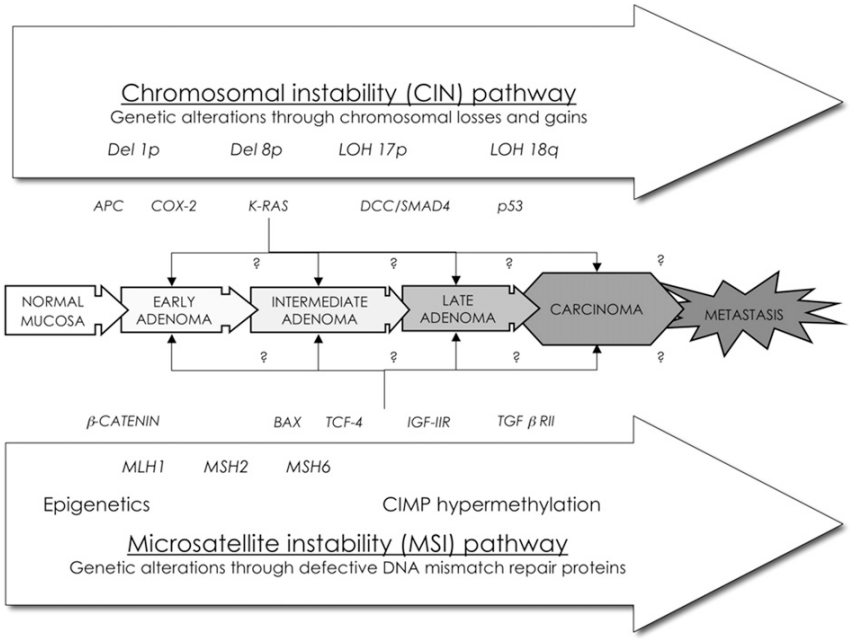


Figure 2.1: Major elements of the Vogelgram. Reproduced from [Soreide et al., 2008]

of β-catenin and may lead to downstream activation of the Tcf/LEF transcription factors.

Alternatively, the β-catenin gene *CTTNB1* may itself become mutated, resulting in a failure to appropriately bind to the APC-mediated stabilising complex, thus leading to cytosolic accumulation, with the same downstream effect of Tcf/LEF activation [Morin et al., 1997, Polakis, 2000, Miller, 2002, Mann et al., 1999]. Mutational surveys suggest that the frequency of *CTNNB1* mutation in sporadic colorectal cancer is approximately 1% [Polakis, 2007].

Experimental evidence demonstrates that even a single wild-type APC allele is sufficient to block the Wnt cascade. Thus, both copies of the gene must exhibit sufficient mutation to result in phenotypic tumour progression [Oshima et al., 1995]. Oshima et al. have suggested a polyp formation hypothesis in APC-mutant mice whereby the pre-malignant polyp is initiated by abnormal cell proliferation at the crypt-villous boundary [Oshima et al., 1997]. This histological evidence suggests that the earliest micro-adenomas originate from the

zone of proliferation and grow in an abnormal direction. Oshima et al. characterise the histology of these micro-adenomas as "abnormal tissue building" and make the important distinction that the tissues do not exhibit a faster growth rate. Oshima et al. further observe that the earliest micro-adenomas are still covered by normal epithelium while advanced tumours are likely to contain cells of mesenchymal origin due to tissue remodelling. These studies indicate that the direct consequence of APC loss is abnormal tissue architecture with an enlarged proliferating crypt compartment [Oshima et al., 1997] (See Figure 2.2). The cascade of genes activated downstream of β-catenin Tcf/LEF signalling are consistent with the tumorigenesis theories of Oshima, including regulators of cellular growth, differentiation, and tissue morphology. Reviews of the relationship of

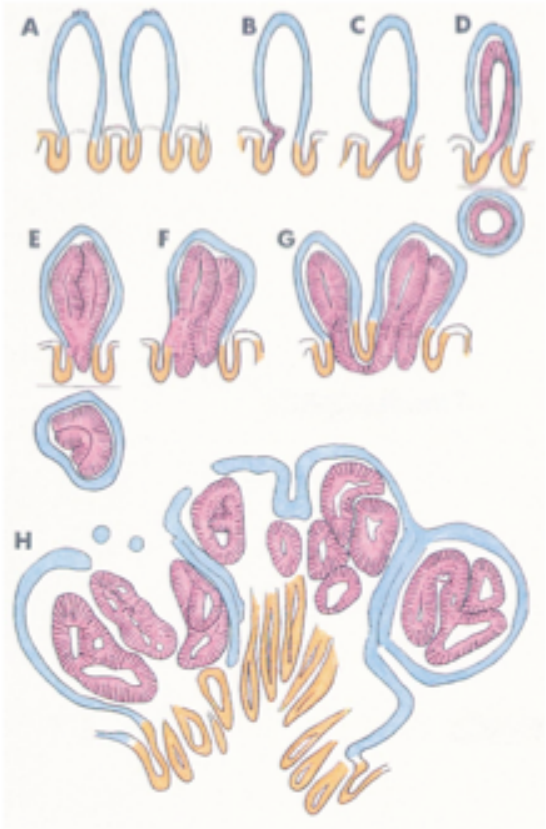


Figure 2.2: Illustration of Oshima's conceptual process of adenoma formation in the small intestine. Reproduced from [Oshima et al., 1995].

colorectal cancer and the Wnt pathway can be found in Kinzler and Vogelstein [1996], Bienz and Clevers [2000], Wong and Pignatelli [2002], and Waterman

[2004].

2.3.4 Chromosomal instability pathway

Colorectal cancer is understood to progress down the adenoma-carcinoma sequence following one of two pathways that lead to the "mutator" phenotype [Soreide et al., 2008, Parsons et al., 1993].

A large number (50%-85%) of colorectal cancers present with altered chromosomal number and/or structure [Lengauer et al., 1997, Narayanan et al., 2003, Goel et al., 2007]. Further, the fact that aneuploidy events have been shown to be consistently associated with loss of function of mitotic checkpoints [Cahill et al., 1998] has led to speculation that chromosomal instability (CIN) represents the most common canonical molecular pathway of oncogenesis [Rasnick and Duesberg, 1999, Soreide et al., 2008].

Ilyas et al. [1999b] note, however, that aneuploidy affects gene dosage and has little or no affect on gene function. Ilyas et al. further argue that while gene dosage can affect cellular function, dominant oncogenes and tumour suppressor genes are more likely affected by altered function (i.e. mutation.) [Ilyas et al., 1999b]. The research by Platzer et al. [2002] demonstrating that there is little correlation between chromosomal duplication and gene expression levels would seem to support this view [Platzer et al., 2002]. On the other hand, an alternative view emphasising that aneuploidy itself may be sufficient to initiate the oncogenic cascade is presented by Rasnick and Duesberg [1999].

Studies which link chromosomal instability with *APC* mutations further emphasise the role of that molecule in epithelial cell regulation within the colonic crypts [Fodde et al., 2001, Kaplan et al., 2001]. By localising to the microtubule "plus" ends, APC has been shown to be a regulator of the cytoskeleton [Mogensen et al., 2002]. Further, while bound to microtubule ends during mitosis APC becomes embedded within kinetochores and complexes with checkpoint proteins Bub1 and Bub2 [Kaplan et al., 2001] Mutated species of APC (i.e. truncated

isoforms) appear to lose this binding ability in vitro and may therefore interfere with normal chromosomal separation during anaphase (reviewed in Narayanan et al. [2003]).

Interestingly, Zumbrunn et al. [2001] also show that GSK3β phosphorylation of APC appears to decrease the association with microtubules – in contrast to APC association with β-catenin which is improved by phosphorylation. This observation lead Zumbrunn et al. [2001] to suggest that APC binding to microtubules and β-catenin may be mutually exclusive events and that GSK3β may act as a molecular switch between the two activities of APC.

2.3.5 The microsatellite instability pathway

The hallmark of the second canonical pathway underlying colorectal oncogenesis is instability within runs of mono- or di-nucleotide repeats called microsatellites [Peltomaki et al., 1993, Aaltonen et al., 1993, Thibodeau et al., 1993]. This microsatellite instability (MSI) occurs when the DNA mismatch repair (MMR) process fails to recognise and correct replication errors made by DNA polymerase during DNA synthesis. The repeat-rich runs of microsatellites render them sensitive to such repair faults.

Human nonpolyposis colorectal cancer (HNPCC) is an autosomal dominant syndrome resulting from mutations to MMR genes and microsatellite stability is often used diagnostically to confirm individuals likely to suffer from this syndrome [Liu et al., 1996]. Families afflicted by HNPCC most commonly exhibit mutations in the two key MMR genes called *MSH2* and *MLH1*, human homologues of the bacterial genes *MutS* and *MutL*, respectively [Peltomaki and de la Chapelle, 1997]. Mutations have also been identified in *MSH6*, *MSH3*, and *PMS2*. Along with *PMS1* these six genes code for error-specific hetero-dimer complexes that recognise and eliminate base-base mismatches and insertion-deletion loops caused by DNA polymerase slippage. There have been more than 400 different mutations identified in these key MMR genes [Peltomaki, 2001]. (Reviewed in Peltomaki [2003], Lucci-Cordisco et al. [2003] and Grady [2004]).

Approximately 10-15% of sporadic colorectal cancers are characterised by microsatellite instability [Boland, 2000, Samowitz et al., 2001]. Unlike the MMR mutations associated with HNPCC, however, sporadic cancers with MSI are almost exclusively the result of *MLH1* gene silencing by methylation [Herman et al., 1998, Toyota et al., 1999, Deng et al., 1999]. While there has been relatively little literature specific to the relationship of MMR defects and pre-malignant lesions, the nature of this phenotype inherently leads to higher rate of accumulated mutation [Parsons et al., 1993] and progressive imbalance of the molecular regulatory mechanisms within the cell [Boland, 2000].

MMR failure is especially likely to affect gene targets with repeating sequences. Among the known affected genes are *TGFBR2*, *IGF2R*, *BAX*, *TCF4*, *PTEN*, *E2F4* and *AXIN2* [Peltomaki, 2001, Souza et al., 1997, Woerner et al., 2003]. In addition to the loss of tumour suppressor activity by these key regulatory genes (*TGFBRII*, *IGF2R*,*BAX*, *E2F4* and *PTEN*), mutations to *TCF4* and *AXIN2* both have the effect of stimulating the Wnt signalling that is described in detail above. Consequently, MMR failure can be linked to pre-malignant progression [Fukushima et al., 2001].

Nevertheless, mutations (including to *APC* or *CTNNB1*) that result in increased proliferation, and consequently polyp formation, precede malignancy [Oshima et al., 1995]. Current theories of oncogenesis and malignancy suggest that progression to colorectal carcinoma necessarily requires further cell disruption or mutations [Kinzler and Vogelstein, 1996].

2.3.6 The methylator phenotype

In addition to the "mutator" phenotype (i.e. tissues exhibiting CIN or MSI), recent evidence has begun to point to a third pathway associated with epigenetic silencing generally, and often methylation of CpG islands in gene promoters in particular [Soreide et al., 2008]. Methylation of these cytosine-guanosine dinucleotide rich sequences inhibits gene transcription [Bird, 1986]. The CpG island methylator phenotype (CIMP) may occur in 20%-40% of sporadic colorectal

cancers although there is uncertainty about whether the CIMP represents a true mechanistic pathway or just an accumulation of random events [Weisenberger et al., 2006, Jass, 2007b, Goel et al., 2007, Ogino and Goel, 2008].

Interestingly, the CIMP positive cancers are characterised by distinct pathological and clinical features, including high frequency of proximal lesions, association with older patients and females, and frequent MSI [Soreide et al., 2008, Goel et al., 2007].

2.3.7 Serrated polyp pathway

Colorectal epithelial polyps are historically divided into two classes: neoplastic adenomas and hyperplastic polyps [Longacre and Fenoglio-Preiser, 1990, Jass, 2005]. Hyperplastic polyps have long been presumed to be benign polyps unrelated to cancer progression[Muto et al., 1975]. There is, however, emerging evidence that hyperplastic polyps belong to a superset of "serrated" polyps, some of which may represent a distinct colorectal cancer pathway that is independent of the traditional adenoma-carcinoma sequence. This pathway is called the "serrated polyp pathway" or the "serrated neoplasia pathway" [Jass, 2005, Liang et al., 2008, Soreide et al., 2008, Hawkins et al., 2002].

The term "serrated adenoma" was coined by Longacre and Fenoglio-Preiser in 1990 to describe a mixed polyp that is morphologically similar (though possibly not molecularly similar) to the hyperplastic polyp but cytologically similar to an adenoma [Longacre and Fenoglio-Preiser, 1990]. While early reports of serrated adenomas suggested a molecular biological profile consistent with the adenoma-carcinoma sequence (e.g. LOH in *APC*, mutations in *KRAS* and *TP53*, etc.) accumulated evidence shows that many serrated adenomas do not reveal these mutations. Serrated adenomas do not exhibit chromosomal instability and they demonstrate stable, wild-type Wnt cascade control. On the other hand, serrated adenomas generally include mutations in either *BRAF* or in rare cases *KRAS* (but not simultaneously), mutations of *TGFβRII*, silencing of *MGMT* and *MLH1*, and elevated levels of methylation [Jass, 2005, 2007b].

There is evidence that sporadic cancers with MSI arise from serrated adenomas and these tumours often show a loss of *hMLH1* [Hawkins and Ward, 2001]. In this serrated polyp pathway, the serrated adenoma appears to be an intermediate form between hyperplastic polyps and adenocarcinoma [Hawkins and Ward, 2001, Jass, 2005, Soreide et al., 2008]. Further most, if not all, tumours characterised as CIMP-high also progress through the serrated polyp pathway [Soreide et al., 2008].

2.3.8 Other Pathways

In addition to the major pathways of colorectal tumorigenesis there is evidence that other pathways may exist. The most well documented of these is the ulcerative colitis associated colorectal carcinomas (UCACC) [Potter, 1999, Ilyas et al., 1999b].

Colorectal cancer in ulcerative colitis patients appears to differ in both presentation and in the molecular oncogenesis following from colitis. For example, in contrast to the polyploid adenomas that are precursors to most sporadic colorectal cancer cases, many UCACC patients present with diffuse, flat adenomas [Ilyas et al., 1999b]. Unfortunately, though, beyond correlations to increased mutations (e.g. TP53) [Fogt et al., 1998] and changes in expression of particular genes (e.g. Bcl-2) [Ilyas et al., 1996], there is little substantive understanding concerning the nature of the molecular mechanisms of this pathway (reviewed in Benhattar and Saraga [1995], Ilyas et al. [1999b] and so it will not be further discussed here.

2.3.9 Acceleration of cancer progression by TGF-β and the Epithelial-Mesenchymal Transition

Despite the role of TGFβ as a strong inhibitor of epithelial proliferation in the normal mucosa, TGFβ signalling has also been shown to accelerate the onset of aggressive carcinoma in cancers in an oncogenic manner. This apparent paradox

appears related to TGFβ's action as an inducer of the essential developmental process called the "epithelial-mesenchymal transition" (EMT). The hallmark of this transition is the shift from an epithelial phenotype characterised by strong cell-to-cell communications and rigid cell polarity to a mesenchymal phenotype that involves weaker cell interactions, increased motility, and the non-polarised fibroblast cell morphology [Bates and Mercurio, 2003]. In vertebrate development the ability to transition from sheets of epithelial cells to mesenchymal cells is fundamental to organogenesis of the heart, musculoskeletal system, the peripheral nervous system as well as most cranial/facial features [Liotta and Kohn, 2001].

In carcinogenesis, TGFβ may stimulate an EMT-like event that confers phenotypic selective advantage on tumour cells and endows such cells with increased metastatic potential. Once the cell escapes the inhibitory epithelial controls, TGFβ signalling may endow the transformed cell with the critical characteristics of escape, invasion, and motility [Bates and Mercurio, 2005].

2.4 Colorectal neoplasia biomarker research

The availability of high-throughput technologies for measuring phenotypic data such as the transcriptome, proteome and epigenome has lead to a rapidly increasing field of biomarker studies [Nannini et al., 2008]. There are now many published reports which explore colorectal gene expression and candidate biomarkers based on differential expression. An overview of this literature, including a list of microarray experiments measuring colorectal tissue specimens, is included as an Appendix (See Appended Chapter A) and the key conclusions of this review are discussed here.

2.4.1 Microarray data for discovery

There are many studies which compare gene expression in neoplastic colorectal with non-neoplastic controls by measuring the concentration of one or more

mRNA transcripts using, for example, microarrays. Differential gene expression between these phenotypes is reported in each of the more than 70 microarray studies reviewed here. Further, there is an emerging agreement between the results of these studies for a number of particular genes [Chan et al., 2008]. These observations provide a foundation of support for the hypothesis examined here, namely that gene expression patterns can be used to discriminate between colorectal phenotypes.

A systematic review of this literature, however, suggests that many reports of colorectal neoplasia discovery are limited by two common weaknesses. The first common weakness is the lack of non-neoplastic diseased specimens in most studies. Colorectal neoplasia is not the only colorectal tissue phenotype which may be correlated with altered gene expression patterns relative to normal tissues. Ulcerative colitis, for example, has been shown to exhibit differentially expressed genes relative to healthy colorectal tissues [Eriksson et al., 2008]. Failure to include non-neoplastic diseased tissues increases the risk of identifying candidate genes which are not specific for neoplasia [Pepe et al., 2001]. Of the literature reviewed here, only the work of Galamb et al. included both healthy normal controls and non-neoplastic diseased controls (in this case, inflammatory bowel disease (IBD)) for comparison to neoplastic specimens. Given the potential that other diseases, including colitis, could affect gene expression patterns in colorectal tissues, this lack of non-neoplastic disease controls is a key weakness of the prior literature.

Another common problem with gene expression biomarker studies reported in the literature is small sample size, generally. Among the microarray experiments measuring colorectal specimens, the largest studied identified here measured 168 specimens (84 cancers and 84 matched normal controls) [Kim et al., 2008a]. Using a limited number of discovery specimens increases the risk that candidate biomarkers will fail to perform well across the full range of tissues defining a particular phenotype. Colorectal neoplasia is increasingly being recognised as a heterogeneous disease whose aetiology may involve multiple possible genomic pathways [Jass, 2007b]. Studies which analyse limited data sets are less likely

to discover biomarkers which will be sensitive for disease across the full range of this heterogeneity.

Further, experiments described herein provide clear evidence that even within non-diseased tissues there is evidence of differential gene expression, such as along the longitudinal axis of the large intestine. Inclusion of a larger sample size provides a measure of protection against bias introduced by such confounding variables if the discovery data cannot be properly balanced by design, or even if sources of confounding variables are not fully recognised.

2.4.2 The need for validation

Despite a large and growing body of biomarker discovery literature, no new biomarker candidate has gained broad acceptance as a marker for colorectal neoplasia [Ransohoff, 2004b]. The lack of a compelling biomarker candidate arising from the literature may be due to the fact that few biomarkers survive subsequent validation studies using independent clinical specimens. On the other hand, microarray validation of selected differentially expressed genes using PCR-based confirmation is relatively common [Canales et al., 2006]. Many of these PCR experiments, however, are carried out using RNA extracts also used for microarray-based discovery. Thus, while these data provide confirmation that transcripts discovered using microarray probesets are likewise detectable or differentially expressed using this alternative technology, these experiments are not evidence of clinical validation using independent tissues.

2.5 Conclusions

In this chapter the literature related to colorectal gene expression was presented with an emphasis on the biological processes of colorectal adenoma development. The organogenesis of the large intestine is broadly patterned on a development program that further differentiates the underlying intestinal phenotype. A key

element of this pattern is the development of complex crypt-surface dynamics that provide balance between the exfoliated epithelial lining of the colorectal lumen and the active stem cell compartment of the crypt. Adenomatous polyps (including serrated adenomas), on the hand, reflect a disequilibrium of these forces giving rise to neoplastic tumours and possibly, in some polyps, *in situ* carcinoma. Finally, an epithelial to mesenchymal tissue transition may prime tumour cells for metastasis and potentiate the invasive cancer phenotype.

The Wnt pathway appears to play a central role in both the development and maintenance of the crypt-axis architecture and in oncogenesis. Control of this pathway is often disrupted early in oncogenesis as described by the adenoma-carcinoma sequence. Consequently, analyses of Wnt-associated genes may provide useful clues about molecular markers for colorectal adenomas.

Colorectal cancer is increasingly recognised as a heterogeneous disease [Jass, 2007b]. The suggestion of the serrated polyp pathway as a possible alternative to the classical adenoma-carcinoma sequence could improve the molecular understanding of colorectal oncogenesis and may to lead to improved clinical management [Jass, 2007b, Ogino and Goel, 2008]. Collectively, CIMP status, MSI, and CIN are emerging as defining variables in the molecular classification of colorectal cancer [Ogino and Goel, 2008, Jass, 2007b,a, Soreide et al., 2008].

2.5.1 Hypothesis in the context of the literature

Based on this review of the colorectal gene expression literature, there is evidence of differentially expressed genes in neoplastic colorectal tissues compared to non-neoplastic controls. The literature does not, however, adequately address the main hypothesis of this thesis: that gene expression biomarkers can be used to accurately discriminate or predict colorectal neoplastic tissues from non-neoplastic controls. Evidence of differentially expressed genes is not, in itself, convincing evidence that genes can be used to predict neoplasia prospectively. To address this hypothesis, this thesis describes research aimed at first discovering and then validating candidate gene expression biomarkers which can be

used to define discriminant rules for classification of colorectal tissue as either neoplastic or non-neoplastic.

Another aspect of this thesis is to identify colorectal neoplasia markers that are sensitive and specific for *both* colorectal adenomas and adenocarcinoma. While biomarkers for precancerous colorectal adenomas are not well studied [Sabates-Bellver et al., 2007], these neoplasms provide the key to prevention of cancer in addition to the reduced mortality achieved by early detection of cancer by screening [Levin et al., 2008]. Interestingly, however, there is evidence for commonly differentially expressed genes in those few studies which test adenomas.

The heterogeneity of colorectal neoplasia through one of several pathways of oncogenesis may pose a challenge to achieving the principal aim of this thesis. The evidence that tumours manifesting MSI yield differential transcription patterns relative to MSS tumours underscores this concern [Soreide et al., 2006]. On the other hand, the high dimensional nature of gene expression microarray technology may provide sufficient phenotype resolution to identify either a single gene biomarker common to all neoplastic phenotypes or else a multi-gene panel that adequately captures the heterogeneity of neoplasia.

Despite the growing number of research papers in this field, there is currently no clinically useful biomarker that is sensitive and specific for both colorectal carcinoma and benign precancerous adenomas. There are, however, many examples of biomarker "fishing expeditions" that claim to have found promising leads [Soreide et al., 2008] and numerous examples of promising discovery research followed by poor validation experience [Ransohoff, 2004b].

This lack of validated biomarkers is addressed by this thesis by exploring the hypothesis that gene expression biomarkers can be used to discriminate colorectal neoplastic tissues from non-neoplastic controls.

Chapter 3

Discriminant Analysis: Pattern Classification with Gene Expression Data

This chapter aims to review the mathematical framework of statistical learning and decision theory as related to the material in this thesis. In particular, classical discriminant techniques such Fisher's linear discriminant analysis [Fisher, 1936], quadratic discriminant analysis and some extensions thereof are reviewed [Rao, 1948]. In subsequent chapters particular attention is given to the situation where the number of features (e.g. genes or probesets) exceeds the number of observations, often referred to as the $p > n$ condition.

3.1 Background

This research aims to analyse gene expression data to discover biomarkers that are useful for the diagnosis of colorectal neoplasia including adenomas and adenocarcinoma. From a mathematical perspective, this objective involves two discrete but intimately related steps. The first step is to learn a discriminant function that distinguishes between (or separates) the phenotypes of interest in the feature space of chosen variables. In supervised learning, this step is

"training" and is performed using data of known classification, for example gene expression levels measured in tissues of known phenotype. Discovering such classifiers is the primary domain of statistical learning theory [Hastie et al., 2001, Vapnik, 1995]. The ultimate goal, however, is not to classify tissues of known phenotype, but rather to predict the phenotype of unclassified tissues[1]. To derive a practical outcome from this analysis one must transform the discovered discriminant function into a classifier rule that can be used to interrogate novel observations in the future and predict class [Hand, 1997, McLachlan, 1992].

The application of statistical learning theory rests on the belief that a discriminant function discovered in a small sample of observations (e.g. tissues) of known class (e.g. disease vs. normal) generally referred to as a "discovery", "design" or "training" set can be applicable to building classification rules for future use [Duda et al., 2000]. This assumption is the essence of supervised learning whereby the design set is said *to supervise* the discriminant discovery process [Ripley, 1996].

Many of the mathematical techniques that are utilised in this research are motivated by a need to extract information from the design or training set that will generalise to the wider population of tissues of a given phenotype which will be investigated in the future. Many statistical learning algorithms are refined to avoid overfitting the discriminant function to the design data [Hand, 1997]. These mathematical techniques, however, will not overcome biological or selection bias that may be present in the training set if the tissues of the design set do not adequately represent the "true" nature of tissues, including the breadth of natural variation, one aims to classify in a biological sense [Ransohoff, 2004b].

More generally one should also consider the assumption that gene expression provides an appropriate representation of classes of interest [Duda et al., 2000]. The work presented in this thesis rests on an a belief that colorectal tissue phenotype can be captured by a vector of gene expression values (continuous real numbers) and that a robust phenotype classifier can be constructed by measuring

[1]Note that in the original treatment Fisher [1936], discriminant analysis involved a more general analytical technique designed to interrogate relationships without regard to classification.

the structural relationships between these data. In other words, there is an assumption that the choice of representation (gene expression) is correlated with intrinsic properties of each phenotype under study. There is reasonable support for this assumption and a full review of gene expression pathways perturbed in colorectal neoplasia is provided in Chapter 2. In general, this assumption is critical because phenotypic class separation is almost always guaranteed in moderate, to high-dimensional gene expression data [Hastie et al., 2001].

Finally, discriminant function discovery involves detecting quantitative relationships in the observational data between the classes of interest in the chosen representation data space. In this work, univariate and multivariate analyses are applied to gene expression data to predict neoplastic status. Clearly, an analysis of such inter-class gene expression differences could be a justifiable, practical aim if simply understanding these differences provides valuable insight and clinical utility generally. For example, any gene expression pattern that is observed raises the biological interpretation question: *Why is this particular gene expression pattern observed and/or changing between phenotypes?* This thesis avoids this perspective except in narrow circumstances. Further, one should be cautious with regard to the biological interpretation of a discriminant function that is discovered on the basis of magnitude of signal difference (even if often with respect to intra-class variation). The mathematical utility of large magnitude signal changes should not be confused with issues of biological relevance such as disease aetiology although possible avenues of biological importance may be indicated in some circumstances. Ultimately, the robustness and generalisability of a particular discriminant model may be well served by gaining a full understanding of the underlying biological perturbations which lead to the gene expression patterns we describe here. Nevertheless, such understanding is not the principal aim of this work which is to accomplish the first step of this process: to identify a robust classifier of neoplastic status using gene expression data. Where appropriate, however, possible avenues of biological impact may be indicated.

3.1.1 Discrimination between two classes

The goal of this work is to identify biomarkers useful for discriminating neoplastic tissues from non-neoplastic "normal" tissues. While we occasionally investigate multiple class relationships beyond the adenoma vs. normal comparison such as extensions to cancer, adenoma staging, and non-neoplastic diseases such as colitis, etc. the principal domain of this analysis involves discriminating between two phenotypic conditions at a time, for example cancer tissues versus normal tissues, adenomatous vs. non-neoplastic tissues, etc. Therefore, this review will generally focus on the two-class discriminant case where, we often benefit from mathematical simplifications.

Some classifiers generalise from the two-class case to more classes in a very natural way. Other classifiers are more intrinsically 2-class and require more elaborate schemes (all pairwise comparisons, etc.) to generalise them to multiple classes. See Hastie and Tibshirani [1998] for discussion and references.

The gene expression data explored here are usually measured using oligonucleotide microarrays. Data pre-processing such as background correction and inter- and intra- experiment normalization are discussed in Chapter 5 and Section 5.5. The problem of microarray normalization is an area of ongoing development [Irizarry et al., 2003]. The background corrected and normalized gene expression data analysed here are assumed to be non-negative, continuous real numbers. Consequently, this mathematics review will be restricted to treatment of the continuous real case without reference to discrete, categorical, or mixed data.

3.2 Statistical decision theory

This thesis aims to discover patterns of biomarker gene expression that will have clinical utility in the medical decision process. The process of rational medical disease diagnosis in the context of colorectal neoplasia can be described in the

formal terms of statistical decision theory. For a review of medical decision making see Spring [2008].

Without loss of generality, this formal treatment of colorectal neoplasia diagnosis and medical decision making is restricted to the two-class exclusive case:

$$\mathbf{C_1}: \quad \text{Class1} \quad \text{Non} - \text{neoplasia}$$
$$\mathbf{C_2}: \quad \text{Class2} \quad \text{Neoplasia}$$

A theoretical decision machine will involve assigning a tissue to one of these two classes by making a choice between one of two diagnostic possibilities:

$$\mathbf{D_1}: \quad \text{Negative diagnosis} \quad \text{Neoplasia determined absent}$$
$$\mathbf{D_2}: \quad \text{Positive diagnosis} \quad \text{Neoplasia determined present.}$$

From a clinical diagnostic perspective the relationship between true class membership, or phenotype, and diagnosis is as follows:

	$\mathbf{D_1}$	$\mathbf{D_2}$
$\mathbf{C_1}$	True Negative	False Positive
$\mathbf{C_2}$	False Negative	True Positive

For completeness, one could extend this framework by indicating that each $\mathbf{D_x}$ will rationally be followed by an action $\mathbf{A_x}$, where e.g. a positive diagnosis ($\mathbf{D_2}$) leads to appropriate clinical follow-up ($\mathbf{A_2}$). However, if we assume the ideal case that a rational action will automatically follow from a diagnosis, we can simply ignore this detail and focus on the diagnosis itself.

3.2.1 The base case: Disease incidence known, no training data

In the simplest analytical case the only knowledge one has available is disease incidence in the population of interest, i.e. the fraction of the population afflicted.

$$P(\mathbf{C_2}) = \frac{\text{total number of neoplasia cases}}{\text{total population size}} = \text{incidence},$$

where P here indicates probability.

Setting aside the practical implications or downstream effects of making a decision $\mathbf{D_1}$ or $\mathbf{D_2}$, the rational decision theoretic approach leads us to simply assign any unknown observation to that group that is more likely based on incidence. In the two-class case (e.g. disease or healthy, exclusive), incidence of healthy and diseased individuals is called the *a priori* probabilities (or just priors), $P(C_1)$ and $P(C_2)$. The trivial decision rule is thus

$$\text{Decide } \mathbf{D_1} \text{ iff } P(\mathbf{C_1})/P(\mathbf{C_2}) \geq 1, \mathbf{D_2} \text{ otherwise.}$$

In the case of $P(\mathbf{C_1}) = P(\mathbf{C_2})$, either choice is equally rational and one must simply choose.

3.2.2 General case: Disease incidence known, data available

In the supervised case (that is the subject of this work) further data is available. This training data, $\mathcal{T}$, usually manifests as a matrix of N observations by p features (e.g. genes or probesets),

$$\mathcal{T} = \underset{N \times p}{\mathbf{X}},$$

where $\mathbf{x}_i$ is a p length (real) vector which describes a single observation (row of $\mathbf{X}$) and $i \in N$.

The supervised learning paradigm is founded in the belief that there is a class-conditional probability density for each phenotype $\mathbf{x}$, $P(\mathbf{x}|\mathbf{C})$ and such conditional densities are separable. This belief is sustainable in this work if expression for selected genes is class/phenotype dependent.

To construct the general decision case involving data, we first note that any joint probability can be factored into the product of a conditional probability and a marginal probability as follows:

$$P(A, B) = P(A|B)P(B).$$

This relation is useful when applied to the joint distribution of our observed data $\mathbf{x}_i$ and $P(\mathbf{C_j})$ (i.e. phenotype incidence). By combining the prior knowledge of class incidence $P(\mathbf{C_j})$ and the class-conditional likelihood of the data $p(\mathbf{x}_i|\mathbf{C_j})$ for our two class case $j \in \{1, 2\}$, we can estimate an *a posteriori* probability (or posterior) that a given observation is a member of $\mathbf{C_j}$. Note that

$$p(\mathbf{x}_i, \mathbf{C_j}) = p(\mathbf{C_j}, \mathbf{x}_i),$$

$$p(\mathbf{x}_i|\mathbf{C_j})P(\mathbf{C_j}) = p(\mathbf{C_j}|\mathbf{x}_i)P(\mathbf{x}_i)$$

where in the two class case $j \in \{1, 2\}$ and

$$p(\mathbf{x}_i) = p(\mathbf{x}_i|\mathbf{C_1})P(\mathbf{C_1}) + p(\mathbf{x}_i|\mathbf{C_2})P(\mathbf{C_2}).$$

Thus, the posterior estimate is derived as

$$p(\mathbf{C_1}|\mathbf{x}_i) = \frac{p(\mathbf{x}_i|\mathbf{C_1})P(\mathbf{C_1})}{p(\mathbf{x}_i|\mathbf{C_1})P(\mathbf{C_1}) + p(\mathbf{x}_i|\mathbf{C_2})P(\mathbf{C_2})}, \text{ and} \tag{3.1}$$

$$p(\mathbf{C_2}|\mathbf{x}_i) = \frac{p(\mathbf{x}_i|\mathbf{C_2})P(\mathbf{C_2})}{p(\mathbf{x}_i|\mathbf{C_1})P(\mathbf{C_1}) + p(\mathbf{x}_i|\mathbf{C_2})P(\mathbf{C_2}).}. \tag{3.2}$$

This is Bayes' theorem [Gelman et al., 2004]. The denominator of the posterior probability, which is the same for both class posterior estimates, acts as a scaling factor to make sure that the posterior probabilities over all $\mathbf{C_j}$ sum to one.

In the context of neoplasia diagnosis, Bayes' theorem estimates the probability that an observation $\mathbf{x}_i$ is neoplastic by adjusting our prior belief of disease incidence by the estimate of the neoplasia likelihood given the data $p(X|\mathbf{C_j})$.

These posterior probabilities for $\mathbf{C_1}$ and $\mathbf{C_2}$ can be compared as above to construct a decision rule based on the data:

$$k = \frac{p(\mathbf{C_1}|X)}{p(\mathbf{C_2}|X)} = \frac{\dfrac{p(X|\mathbf{C_1})p(\mathbf{C_1})}{p(\mathbf{x}_i|\mathbf{C_1})P(\mathbf{C_1}) + p(\mathbf{x}_i|\mathbf{C_2})P(\mathbf{C_2})}}{\dfrac{p(X|\mathbf{C_2})p(\mathbf{C_2})}{p(\mathbf{x}_i|\mathbf{C_1})P(\mathbf{C_1}) + p(\mathbf{x}_i|\mathbf{C_2})P(\mathbf{C_2})}}, \tag{3.3}$$

$$= \frac{p(X|\mathbf{C_1})p(\mathbf{C_1})}{p(X|\mathbf{C_2})p(\mathbf{C_2})}, \tag{3.4}$$

Decide $\mathbf{D_1}$, iff $k \geq 1$; $\mathbf{D_2}$ otherwise.

This decision rule form is called the "likelihood ratio" or the Neyman-Pearson lemma for hypothesis testing [Neyman and Pearson, 1932, Hand, 1997].

3.2.3 Cost and risk Functionals

Finally, this formulation of Bayes' rule makes no distinction for the impact of misdiagnosis in the presence or absence of disease. For many medical diagnostic decisions the costs (by many metrics) of not reporting a positive disease diagnosis for a patient with disease (i.e. false negative) is often not equal to the cost of over-diagnosis (i.e. false positive) [Pepe et al., 2001]. In decision theoretic terms, such cost terms (also called *risk*) and can be introduced using a loss function such as $\lambda(\mathbf{D_j}, \mathbf{C_k})$, the loss function associated with making decision $\mathbf{D_j}$ (perhaps including the cost associated with action $\mathbf{A_j}$) when the true class state is $\mathbf{C_k}$ [Duda et al., 2000].

Thus we can transform Bayes' rule from choosing the class with a maximum posterior probability into a Bayes' risk which attempts to minimise our cost risk associated with a given diagnosis. This cost is called conditional risk when we condition our decision over the data, $\mathcal{R}(\mathbf{D_j}|\mathbf{x}_i)$ [Hand, 1997]. For the two class case we have:

$$\mathcal{R}(\mathbf{D_j}|\mathbf{x}_i) = \sum_{k=1}^{2} \lambda(\mathbf{D_j}, \mathbf{C_k}) P(X|\mathbf{C_k}) \text{ for } j \in \{1, 2\}$$

and we select j to minimise the risk $\mathcal{R}$. For clarity, we expand as follows

$$\begin{aligned} \mathcal{R}(\mathbf{D_1}|\mathbf{x}_i) &= \lambda(\mathbf{D_1}, \mathbf{C_1})P(X|\mathbf{C_1})P(\mathbf{C_1}) + \lambda(\mathbf{D_1}, \mathbf{C_2})P(X|\mathbf{C_2})P(\mathbf{C_2}) \\ \mathcal{R}(\mathbf{D_2}|\mathbf{x}_i) &= \lambda(\mathbf{D_2}, \mathbf{C_1})P(X|\mathbf{C_1})P(\mathbf{C_1}) + \lambda(\mathbf{D_2}, \mathbf{C_2})P(X|\mathbf{C_2})P(\mathbf{C_2}), \end{aligned}$$

and

Decide $\mathbf{D_1}$ iff $\mathcal{R}(\mathbf{D_1}|\mathbf{x}_i) \leq \mathcal{R}(\mathbf{D_2}|\mathbf{x}_i)$ and $\mathbf{D_2}$ otherwise.

Finally, in the simplified case where we assume that there is no cost for making a correct diagnosis, we can simplify Bayes' decision rule further:

Decide $\mathbf{D_1}$ iff

$$\lambda(\mathbf{D_1}, \mathbf{C_1})P(X|\mathbf{C_1})P(\mathbf{C_1}) \leq \lambda(\mathbf{D_1}, \mathbf{C_2})P(X|\mathbf{C_2})P(\mathbf{C_2})$$

and $\mathbf{D_2}$ otherwise.

This application of Bayes' rule provides a general framework for making rational medical decisions given a prior knowledge of disease incidence, new observations in a training set, and costs of misdiagnosis. In practice, this likelihood-based decision rule is often derived by applying one or more discriminant analysis techniques. The following sections detail the most widely used techniques. See also Hastie et al. [2001], Hand [1997], Duda et al. [2000] and Krzanowski and Marriott [1995].

3.3 Discriminant functions

The classification problem for discriminating tissue phenotypes of interest based on gene expression data is restated here. A typical gene expression experiment such as oligonucleotide microarray collects expression data for p genes (a.k.a. features) in N clinical specimens (e.g. tissue samples) that can be expressed in an $N \times p$ matrix $\mathbf{X}$. Each ith row is thus a p-length vector, $\mathbf{x}_i$, containing the expression levels for a given specimen across p genes, where $i \in \{1, \ldots, N\}$.

In the case of supervised discovery, each row of $\mathbf{x}_i$ is a single observation tissue whose class/phenotype is presumed known for training purposes. We represent the N class labels as $\mathbf{y}$, an N-length vector containing class assignment values for each sample. For a two class problem, e.g. neoplasia vs. non-neoplasia tissues, y_i is typically defined by a binary classification scheme e.g. $y_i \in \{0, 1\}$ although any class assignment values may be used.

Our problem, then is to identify some function f that models the expected output $\mathbf{y}$ from the input data $\mathbf{X}$, i.e. find

$$f(\mathbf{X}) = \mathbf{y}$$
$$f(\mathbf{x}_i) = y_i \text{ for } i \in \{1, \ldots, N\}.$$

By comparing the estimated function f with a threshold, we construct a classification rule [Hand, 1997]. Once this classification rule is constructed, we can then apply this model to future data to classify genuinely unknown specimens. Geometrically, if we assume a threshold based on a linear midpoint rule (between $\mathbf{y}_1$ and $\mathbf{y}_2$) with no weight or cost bias in the decision, this assignment is often to the nearer class in p-dimensional space. Alternatively, if the cost functions for misdiagnosis are not equal or the inputs $\mathbf{x}_i$ are treated with unequal importance we compare $f(\mathbf{x}_i)$ with a different threshold to determine classification. In either case the threshold space where $f(\mathbf{C}_1|\mathbf{x}_i) = f(\mathbf{C}_2|\mathbf{x}_i)$ is called the decision surface [Hand, 1997].

3.3.1 Distance metrics for class separation

A useful approach to constructing $f(\mathbf{x}_i)$ is to identify a projection of the data $\mathbf{w}^t\mathbf{x}$ that yields maximum inter-class separation.

However, unlike individual points in (Euclidean) space which can be easily evaluated relative to each other, there are many choices for distance metrics between a set of observations taken as a group. We can, for instance, measure the distance between $\mathbf{x}_{j\in\mathbf{C}_1}$ and $\mathbf{x}_{j\in\mathbf{C}_2}$ (simplified as $\mathbf{x}_1$ and $\mathbf{x}_2$) by comparing the class centroids given by

$$\hat{\mu}_j = \frac{\sum_{k=1}^{N_j} \mathbf{x}_{j_k}}{N_j}, j \in \{1, 2\}. \tag{3.5}$$

In terms of an optimal projection of the data $\mathbf{w}^t\mathbf{x}$, our aim is then to identify $\mathbf{w}$, such that the absolute value of inter-class separation given by

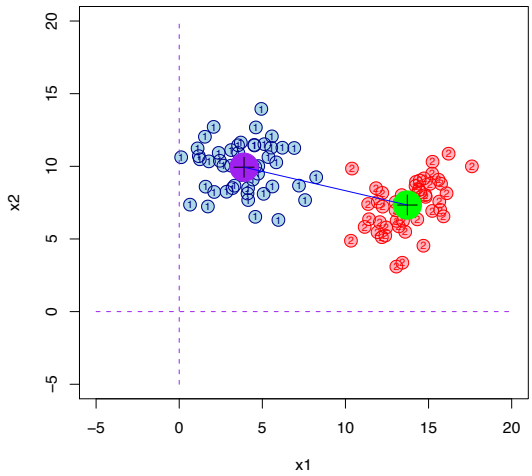
$$D(\mathbf{x}) = |(\mathbf{w}^t\mathbf{x}_1 - \mathbf{w}^t\mathbf{x}_2)|$$

is maximum. To simplify evaluation we can ensure strictly non-negative terms by squaring the terms,

$$D(\mathbf{x}, \mathbf{w}) = (\mathbf{w}^t\mathbf{x}_1 - \mathbf{w}^t\mathbf{x}_2)^2. \tag{3.6}$$

To maximise the distance function, we differentiate Equation 3.6 w.r.t. $\mathbf{w}$. In this case we find that the derivative is constant and the projection simply leads to the line between the two class centroids. This maximum centroid distance is illustrated below in Figure 3.1.

Figure 3.1: Example of centroid-based distance metric where $\overline{AB}$ joins the mean observations in each class)



The Euclidean distance between centroids given in Equation 3.6 is equivalent to the Mahalonobis distance calculated with the assumption of equal variance for each $\mathbf{x}_1$ and $\mathbf{x}_2$. The general form of this distance metric is given by Equation 3.7.

$$D(\mathbf{x})_{Mahalonobis} = \sqrt{(\mathbf{x}_1 - \mathbf{w}^t\mathbf{x}_2)^t\Sigma^{-1}(\mathbf{x}_1 - \mathbf{w}^t\mathbf{x}_2)}. \tag{3.7}$$

Alternatively, assuming that the classes are completely separable, we could measure the distance using the point in $\mathbf{C}_1$ that is most near to a point in $\mathbf{C}_2$ as shown in 3.2. This metric is referred to as the *maximum margin* with respect to individual observations and will be revisited in detail in Section C.1, p.244.

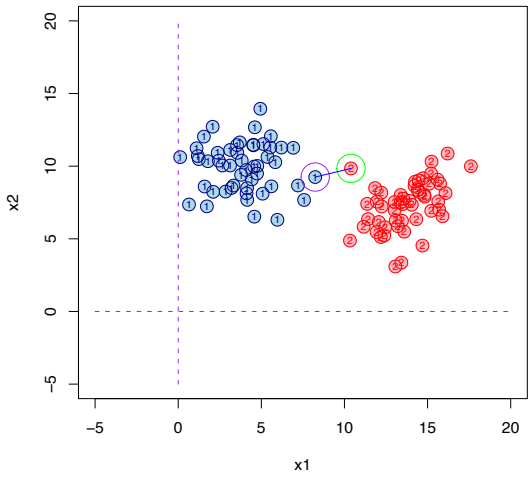


Figure 3.2: Maximum margin separation with respect to the individual points between each class; here $\overline{AB}$ joins the inter-class points nearest to each other.

With the exception of the Mahalonobis metric, these approaches fail to account for class dependent variances [Hand, 1997]. Meta-observations such as the mean vector (centroid) make no allowance for covariance between measurements. One simple solution, therefore, is to standardise inter-class distance metrics using the variances and covariance of the observed training data presented for each class.

In fact, this solution also offers a number of alternative approaches suggested by the choice of the covariance structure(s) that can be used. This choice spans a continuum of complexity where in the most simple case we adopt a common covariance matrix for training data, such as $\mathcal{T} \in \{\mathbf{x}_1, \mathbf{x}_2\}$ while a more complex approach is to calculate separate covariance matrices for each class. Further, a number of regularization techniques can be applied to the later case to shift the covariance matrix back toward some common structure [Krzanowski and Marriott, 1995]. In the extreme case the covariance may be regularised to the identity, $\mathbf{I}$, which brings us back to the centroid solution described above. Using the common (pooled) covariance matrix yields the standard linear discriminant analysis (LDA) rule of Fisher while estimating unique covariance structures for each class results in the quadratic discriminant analysis (QDA) rule [Hand, 1997].

These alternatives are discussed below.

3.3.2 Linear discriminant analysis

We begin by exploring the effect of normalizing the between class distance metric using a common "pooled" covariance structure, such as

$$\hat{\Sigma} = \frac{1}{N_1 + N_2 - 2} \sum_{j=1}^{2} \sum_{i=1}^{N_j} (\mathbf{x}_{ij} - \hat{\mu}_j)(\mathbf{x}_{ij} - \hat{\mu}_j)^t. \tag{3.8}$$

As previously discussed, we restrict ourselves to the two class case. The inter-class distance, D, between $\mathbf{x}_1$ and $\mathbf{x}_2$ projected onto some $\mathbf{w}$ can thus be standardised using $\hat{\Sigma}$,

$$D(\mathbf{x}) = \frac{(\mathbf{w}^t\mathbf{x}_1 - \mathbf{w}^t\mathbf{x}_2)^2}{\mathbf{w}^t\hat{\Sigma}\mathbf{w}}. \tag{3.9}$$

(NB: We again square the numerator for convenience. As this is a monotonically increasing function, this has no affect on the solution.)

Thus D measures a covariance standardized distance between the two classes along the direction defined by $\mathbf{w}$. To find the maximum separation, we find the $\mathbf{w}$ which maximises D. Differentiation of D w.r.t. $\mathbf{w}$ yields

$$\begin{aligned} \mathbf{w} &= c\hat{\Sigma}^{-1}(\mathbf{x}_1 - \mathbf{x}_2) \\ &\propto \hat{\Sigma}^{-1}(\mathbf{x}_1 - \mathbf{x}_2), \\ &\text{ascisarbitrary.} \end{aligned}$$

This projection, which is the basis of linear discriminant analysis (LDA), is shown in Figure 3.3 with the decision surface, which is orthogonal to $\mathbf{w}$.

Intuitively, this approach seems compelling because we can also view maximizing D of Equation 3.9 as finding that linear projection of the data with the largest ratio of the within-class scatter found in the numerator term $(\mathbf{w}^t\mathbf{x}_{j\in\mathbf{C}_1} - \mathbf{w}^t\mathbf{x}_{j\in\mathbf{C}_2})^2$ relative to the between-class variance in the denominator $(\mathbf{w}^t\hat{\Sigma}\mathbf{w})$.

While no distributional assumptions have been made, it is important to note that we have made the choice that each class ($\mathbf{x}_1$ and $\mathbf{x}_2$) can be precisely described in

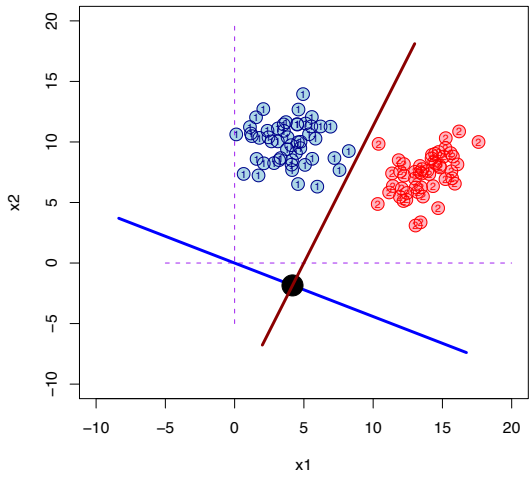


Figure 3.3: Example of Fisher's discriminate function determined using linear discriminant analysis

terms of the class mean ($\hat{\mu}_j$) and a common pooled covariance ($\hat{\Sigma}$). Samples that can be so described (i.e. by the first two moments) are characterised as being drawn from an elliptical distribution (including the multivariate normal) [Hand, 1997]. Thus, while we have not assumed that each class take a multivariate normal form, we have found the solution $\mathbf{w}$ that is optimal for data which are precisely described by the first two moments such as the normal and multivariate normal distribution [Hand, 1997].

In fact, if we make an explicit assumption that each class in $\mathbf{x}_j$ is drawn from a multivariate normal population of p dimensions, we derive a maximum likelihood decision rule equivalent to the LDA solution. Suppose that we model the training data for each class as a multivariate normal with a known $\hat{\mu}_j$ for each class and that both classes share a pooled covariance matrix $\hat{\Sigma}$, we can estimate the probability density function,

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\hat{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\hat{\mu}_j)^t \hat{\Sigma}^{-1}(\mathbf{x}-\hat{\mu}_j)}. \tag{3.10}$$

Assuming that the class prior probabilities are equal, the discriminant function

then is found by comparing the class-conditional densities and assigning future observations to the higher probability. Equivalently we can calculate the log-ratio

$$log\frac{f(\mathbf{x}|j=1)}{f(\mathbf{x}|j=2)} \tag{3.11}$$

and assign to $\mathbf{C}_1$ if the ratio is greater than 1. Substituting from Equation 3.10 above,

$$log\frac{f(\mathbf{x}|j=1)}{f(\mathbf{x}|j=2)} = -\frac{1}{2}(\mathbf{x}_1-\mathbf{x}_2)^t\hat{\Sigma}^{-1}(\mathbf{x}_1-\mathbf{x}_2) + \mathbf{x}^t\hat{\Sigma}^{-1}(\mathbf{x}_1-\mathbf{x}_2). \tag{3.12}$$

If knowledge about prior class probabilities (e.g. incidence) is available, this can also be added to the log-odds ratio,

$$log\frac{f(\mathbf{x}|j=1)}{f(\mathbf{x}|j=2)} = log\Big(\frac{P(\mathbf{C}_1)}{P(\mathbf{C}_2)}\Big) - \frac{1}{2}(\mathbf{x}_1-\mathbf{x}_2)^t\hat{\Sigma}^{-1}(\mathbf{x}_1-\mathbf{x}_2) + \mathbf{x}^t\hat{\Sigma}^{-1}(\mathbf{x}_1-\mathbf{x}_2). \tag{3.13}$$

3.3.3 Least squares (regression) solution

There is a close relationship between LDA and least-squares regression. In the above treatment we have motivated an LDA solution by identifying the hyper-plane by which the inter-class data are best separated. An alternative motivation can be derived by attempting to find the projection vector for which the sum of squared errors between the data and the resultant decision surface is minimum.

In fact, Fisher's original presentation of linear discriminant analysis provides an equivalent regression solution [Fisher, 1936]. As this methodology can provide a convenient link to other methodologies (e.g. to regularised forms involving penalty terms) this derivation is given here.

To pose the regression problem, we first code the class membership of each observation into a target $\mathbf{y}$, such as

$$\begin{aligned} \mathbf{y}_{j=1} &= \frac{N_2}{(N_1+N_2)}, \text{ and} \\ \mathbf{y}_{j=2} &= \frac{-N_1}{(N_1+N_2)} \end{aligned}$$

where N_1 and N_2 are the class member sizes of $\mathbf{C}_1$ and $\mathbf{C}_2$, respectively.

Our goal then is to estimate the regression function $f(\mathbf{X}) = \mathbf{y}$. We accomplish this by seeking a linear combination of the observations ($\mathbf{X}^t\mathbf{w}$) that minimises the difference between the true f and the regression function. This is accomplished in the usual way by estimating $\mathbf{w}$ to minimise the residual sum of squares,

$$RSS = g(\mathbf{w}) = (\mathbf{y} - \mathbf{X}\mathbf{w})^t(\mathbf{y} - \mathbf{X}\mathbf{w}).$$

Differentiating g w.r.t. $\mathbf{w}$, and setting the first[2] derivative to zero (to minimise the function) yields the normal equations:

$$\begin{aligned} \frac{dg}{d\mathbf{w}} &= -2\mathbf{X}^t(\mathbf{y} - \mathbf{X}\mathbf{w}) = 0 \\ \mathbf{w} &= (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}. \end{aligned}$$

Thus,

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{w} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}.$$

For low dimensional problems and where N is sufficiently large, the inverse matrix solution can be estimated using standard techniques such as Gram-Schmidt successive orthogonalization [Golub and Van Loan, 1996]. As p grows, however, algorithms have been shown to suffer from numerical instability and a factorisation method such as **QR** decomposition is preferred [Nakos and Joyner, 1998]. One solves $\mathbf{w}$ using the **QR** decomposition as follows:

[2]We note that it is also desirable to evaluate the second derivative of g to test that all values are non-negative in order to ensure that we have global minimum.

$$\begin{aligned}
\mathbf{w} &= (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y} \\
&= ((\mathbf{QR})^\mathbf{t}\mathbf{QR})^{-\mathbf{1}}\mathbf{X}^\mathbf{t}\mathbf{y} \\
&= (\mathbf{R}^\mathbf{t}\mathbf{Q}^\mathbf{t}\mathbf{QR})^{-\mathbf{1}}\mathbf{X}^\mathbf{t}\mathbf{y} \\
&= (\mathbf{R}^\mathbf{t}\mathbf{R})^{-\mathbf{1}}\mathbf{X}^\mathbf{t}\mathbf{y} \\
&= \mathbf{R}^{-\mathbf{1}}(\mathbf{R}^\mathbf{t})^{-\mathbf{1}}\mathbf{R}^\mathbf{t}\mathbf{Q}^\mathbf{t}\mathbf{y} \\
&= \mathbf{R}^{-\mathbf{1}}\mathbf{Q}^\mathbf{t}\mathbf{y}.
\end{aligned}$$

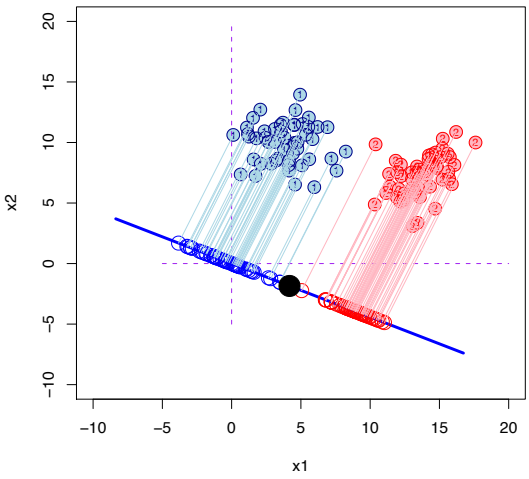
In this case, the estimated solution is then

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{R}^{-1}Q^t\mathbf{y}. \tag{3.14}$$

To apply this solution to future individuals, new observations are projected onto the linear discriminant for comparison to some threshold. In the naive case, both prior probabilities of class membership and mis-classification costs are considered to be equal and one simply assigns new individuals to that class with the closest mean when projected onto $\mathbf{w}$.

For illustration, an application of linear discriminant analysis is shown in Figure 3.4.

Figure 3.4: Example of LDA applied to the same data set from earlier examples.



3.3.4 Quadratic discriminant analysis

The LDA method, as mentioned above, requires an explicit assumption that all classes share a common (pooled) covariance matrix. On reflection, one may consider that this is a rather unlikely circumstance [Hand, 1997]. In fact, we may be more inclined to assume that for any given class, the covariance which describes the differences in relationships between the variables are likely to be significant and relevant. With an increase in calculation complexity we can relax the requirement that there is a common covariance matrix and we derive a quadratic rule,

$$\frac{f(j=2|\mathbf{x})}{f(j=1|\mathbf{x})} = \frac{\frac{1}{(2\pi)^{p/2}|\Sigma_{j_2}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\hat{\mu}_j)^t \Sigma_{j_2}^{-1}(\mathbf{x}-\hat{\mu}_j)}}{\frac{1}{(2\pi)^{p/2}|\Sigma_{j_1}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\hat{\mu}_j)^t \Sigma_{j_1}^{-1}(\mathbf{x}-\hat{\mu}_j)}}, \tag{3.15}$$

which simplifies to:

$$\begin{aligned} = \quad & \mathbf{x}^t(\Sigma_{j_2}^{-1}\hat{\mu}_{j_2} - \Sigma_{j_1}^{-1}\hat{\mu}_{j_1}) - \frac{1}{2}\mathbf{x}^t(\Sigma_{j_2}^{-1} - \Sigma_{j_1}^{-1})\mathbf{x} + log\Big(\frac{\pi_{j_2}}{\pi_{j_1}}\Big) + \\ & \frac{1}{2} + log\Big(\frac{\Sigma_{j_2}}{\Sigma_{j_1}}\Big) - \frac{1}{2}\hat{\mu}_{j_2}\Sigma_{j_2}^{-1}\hat{\mu}_{j_2} + \frac{1}{2}\hat{\mu}_{j_1}\Sigma_{j_1}^{-1}\hat{\mu}_{j_1}, \end{aligned}$$

with the last four terms independent of observation $\mathbf{x}$, and therefore contributing to classification threshold values only.

A comparison of the linear and quadratic decision surfaces for simulated data is shown below in Figure 3.5. In fact, these simulated data were generated using a higher variance for $\mathbf{C}_2$ (in red) which is clearly captured by the quadratic rule. It is, however, important to recognise that while the decision surface is quadratic in the variable space, the solution is still linear in the terms of the model [Hand, 1997].

3.3.5 Overfitting and the bias vs. variance trade-off

The higher complexity decision surface of QDA obviously requires estimation of more parameters than LDA. In many cases data availability is limited and insuf-

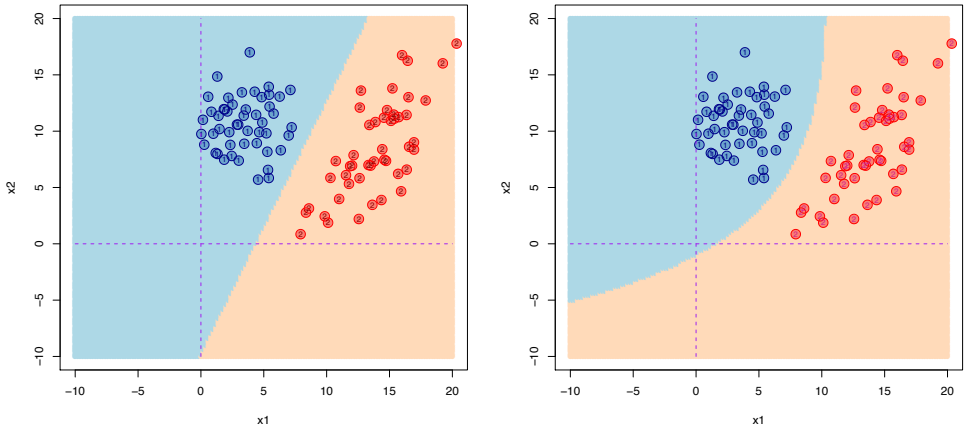


Figure 3.5: Comparison of LDA and QDA solutions against a simulated data set. Note that the red population has a slightly higher variability than the blue population

ficient to accurately estimate the model parameters leading to the possibility of overfitting [Hastie et al., 2001]. In general, overfitting occurs when the solution begins to fit the (random) noise in the data which is specific to the particular observations of that training set instead of the of the underlying class-conditional signal. While overfitting to a design set will yield an apparently accurate model with respect to the training data, such models will not generalise well and future observations will be misclassified because the model parameters poorly reflect the true underlying class-conditional density distributions.

The balance between accurately fitting the training data during supervised learning while likewise attempting to ensure generalisability, or the expected prediction error in *all* future data sets, is at the heart of pattern recognition and carries philosophical as well as technical implications. The fundamental recognition that our goal is not to recognise patterns in the training set but rather to recognise patterns in the data class of interest is vital to this endeavour. Nevertheless, the impossibility of knowing the true underlying class-conditional multivariate distribution offers no alternative but to attempt to discover this pattern from a limited set of training observations. In statistical terms this difficulty also leads to the bias vs. variance components that make up our error.

Given that the data is generated by

$$\mathbf{y} = f(X) + \epsilon,$$

where the expected error mean $E(\epsilon) = 0$ and $\text{var}(\epsilon) = \sigma^2$, the expected prediction error (EPE) for future observations Z based on our training model $\hat{f}(X)$ can be decomposed such that

$$\begin{aligned} EPE(Z) &= E[(\mathbf{y} - \hat{f}(Z))]^2, \\ &= E[f(Z) - \hat{f}(Z)]^2, \\ &= \sigma^2 + [\text{bias}^2(\hat{f}(Z))^2 + \text{var}(\hat{f}(Z))], \\ &= \sigma^2 + (E[\hat{f}(Z)] - f(Z))^2 + E[\hat{f}(Z) - E[\hat{f}(Z)]]^2 \end{aligned}$$

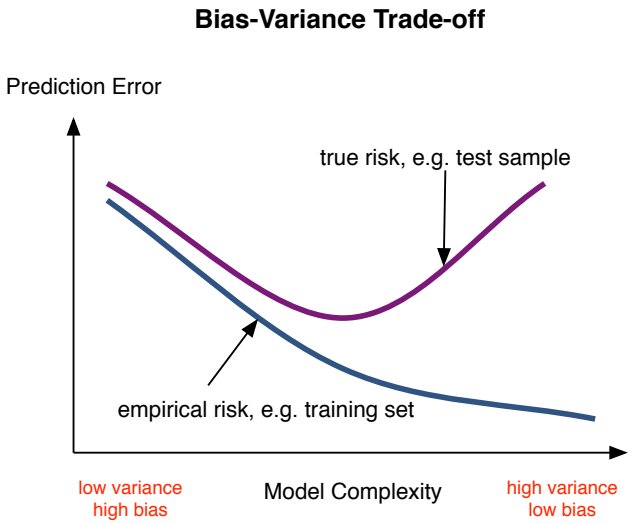
While the irreducible error (σ^2) is data dependent and outside algorithm control, both the bias and variance terms are model selection dependent. These terms combine to form the mean squared error (MSE). The first term of the decomposition is the squared bias which measures the squared difference between the true mean ($f(Z)$) and the expected value estimate averaged over the training data randomness, $E[\hat{f}(Z)]$. The variance term is the variability in the expected estimate.

In general, as the model complexity (i.e. number of parameters) increases, the apparent model fit to the training data will improve and the model variance will increase while the squared bias will decrease. Conversely, decreasing model complexity will decrease model variance and increase bias. Figure 3.6 illustrates the typical bias-variance trade-off relationship to training and generalised model error [Hastie et al., 2001, Nakos and Joyner, 1998].

D. Hand addresses the problem of overfitting with five approaches [Hand, 1997]:

1. Accept a highly flexible model (high complexity) with low bias and attempt to lower the model variance by increasing the number of observations in the training data set. Unfortunately, this approach is often impractical due to limited data availability.

Figure 3.6: Illustration of the bias-variance trade-off. Adapted from Hastie et al. [2001]



2. Improve the selection of p features by e.g. forward or backward feature selection. This approach attempts to lower variance by increasing the bias. However feature selection methods and algorithms are often not obvious and difficult to apply in practice.

3. Constrain model complexity by introducing penalty terms, i.e. regularization. Examples of regularization methods include reducing the data likelihood by a multiple of the number of parameters (Akaike's method, Schwarz's criterion), penalizing the sum of squared parameter terms (ridge regression) and penalizing the absolute value of parameter terms (lasso). These methods attempt to reduce the theoretical MSE (total estimated model error) by increasing bias by an amount less than the corresponding decrease in variance. Regularization and penalization are treated in detail in the following chapters.

4. Smoothing the overfitted function by methods that average or aggregate models including, e.g. bagging.

5. Finally, by attempting to smooth the data itself. As the fundamental difficulty of overfitting is modelling the random variation of the data instead of the underlying function, one can lower the variance of the training data by creat-

ing multiple copies of the data- each copy slightly perturbed from the original data values. Intuitively, this method aims to reduce the noisy "peaks" within the data. Training with noise has been shown to be equivalent to the class of generalised Tikhonov regularisers including, e.g. ridge regression [Bishop, 1994].

These approaches attempt to address the model complexity aspects of overfitting that can occur when the class-conditional signals in the training set are influenced by random noise. However, in the domain of gene expression measurement systems such as oligonucleotide microarray there are far more severe concerns related to model fitting, namely that data features (genes) are likely to be correlated and are often in vast excess of the number of tissues.

The case where the number of features p is much greater than N ($p >> N$) and analytical techniques for dealing with this case will be discussed in the in the following chapter.

3.4 Conclusions

This chapter reviews a formal structure for statistical decision making and introduces the foundations of discriminate analysis including Fisher's linear discriminate analysis (LDA) and the extension to quadratic forms. The effect of cost and risk functionals on decision boundaries was discussed. Finally, the bias-variance trade-off was introduced with the observation that model performance can sometimes be improved by introducing limited bias into model discovery at the expense of higher variance.

The concepts of statistical decision making apply generally to all discriminant analysis problems and the questions of bias-variance trade-off and decision costs are relevant in most real world applications, including clinical decision making using biomarkers. For instance the challenge of establishing a threshold for a positive diagnosis for a particular biomarker measured in a patient specimen is a common problem faced by health care decision makers. In this respect, the material of this chapter is useful for establishing a framework for future

efforts to construct discriminant rules using the validated biomarker results of this research in a real-world clinical context.

The challenge of biomarker discovery which is the aim of this thesis in fact precedes some aspects of discriminant analysis described here. The initial challenge (and the focus of this research) is not to establish a threshold cutoff for positivity for a particular biomarker but rather to discover the biomarker itself! It should be understood that in this context a "biomarker" includes either a single biological molecule or a panel of such molecules, measurements of which are combined to yield a single discriminant score.

With this framework in mind, the discovery data of this research is considered for analysis. Unfortunately, this analysis is immediately confronted with a significant problem: in the case of many genomic-era tools, including gene expression analysis, the number of variables often vastly exceeds the number of patient observations. In this case, it is impossible to analytically determine the inverse of the data covariance matrix. Consequently, the fundamental discriminate analysis tool introduced in this chapter, Fisher's LDA, is immediately rendered useless. This situation is certainly true of the data sets analysed here.

In the next chapter the challenge of discriminant analysis in high dimensional data is considered.

Chapter 4

Discriminant analysis in high dimensional data

4.1 Aims

The preceding chapter introduces discriminant techniques including Fisher's (linear) discriminant analysis. These standard discriminant techniques do not have unique solutions in the case of ill-conditioned or reduced-rank data such as when the number of variables exceeds the number of observations. The aim of this chapter is to review methodologies which may be used to address this difficulty.

4.2 Analysing data with more features than observations

Numerical solutions to discriminant analysis problems using methods such as Fisher's linear discriminant analysis involve solving a linear system to estimate a p vector of coefficients, $\hat{\mathbf{w}}$, in the form of

$$\hat{\mathbf{w}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y} \tag{4.1}$$

However, in many real world applications, including gene expression analysis in particular, the number of features p exceeds the number of observations N, in some cases by orders of magnitude. In these cases, $\mathbf{X}^{\mathrm{t}}\mathbf{X}$ is of reduced rank. Singular matrix products will also arise if $\mathbf{X}^{\mathrm{t}}\mathbf{X}$ exhibits multicollinearity such that some features are exact or near linear combinations of each other [Faraway, 2004]. In geometric terms the problem is characterised by set $\mathcal{T}$ occupying a low-dimensional subspace of the possible feature space $\mathbb{R}^p$ [Hand, 1997].

Setting $\mathbf{A}$=$\mathbf{X}^{\mathrm{t}}\mathbf{X}$, the rank of $\mathbf{A}$ can be explored by constructing the singular value decomposition of $\mathbf{A}$ such that

$$\mathbf{A}= \mathbf{U}\mathbf{\Lambda}\mathbf{V}^{\mathrm{t}} \tag{4.2}$$

where $\underset{N\times N}{\mathbf{U}}$ and $\underset{N\times p}{\mathbf{V}}$ are orthonormal and span the row and column-space of $\mathbf{A}$, and $\mathbf{\Lambda}$ is an N by p diagonal matrix and $\lambda_{max} = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m = \lambda_{min} > 0$ are the m singular values (and in this case, the squared eigenvalues) in decreasing order [Golub and Van Loan, 1996]. Note that λ_{min} is an arbitrary cut-off which is often dependent on machine precision or implementation. Using this decomposition, the rank of $\mathbf{A}$ is given by

$$rank(\mathbf{A})= \max(\mathrm{m}),\ \lambda_m > 0.$$

Thus, in the case where $p \gg N$, the product $\mathbf{X}^{\mathrm{t}}\mathbf{X}$ is not full rank and is not invertible [Nakos and Joyner, 1998]. Hoerl and Kennard note that the uncertainty of (and therefore the expected error) for the estimate $\hat{\mathbf{w}}$ (Equation 4.1) will increase as $\mathbf{A}$ moves from a unit matrix to an ill-conditioned one [Hoerl and Kennard, 1970]. Furthermore, the expected value of the squared distance between the true value $\mathbf{w}^*$ and the estimate $\hat{\mathbf{w}}$ will increase inversely to λ_{min},

$$E\left[(\hat{\mathbf{w}} - \mathbf{w}^*)^t(\hat{\mathbf{w}} - \mathbf{w}^*)\right] = \sigma^2 \sum_{i=1}^{p}(1/\lambda_i).$$

To address this difficulty, Hand reviews four possible approaches [Hand, 1997]:

1. Postulate a particular structure for the covariance matrix to reduce the number of parameters by describing some features as functions of other features. For further details on this approach see Kiiveri [1992].

2. Reduce the number of features by either feature subset selection or by feature extraction. Feature selection methods choose a subset of $k < p$ features for analysis. The subset size k can be established in a forward stepwise fashion starting from zero features, a backward selection process by reducing k stepwise from p, or some combination of the two approaches until some maximum error threshold is achieved. If p is small all-subsets analysis may also be possible.

   Feature extraction involves creating a new set of k features (where again $k < p$) that are a function of, or derived from, the original p-dimensional data. Feature extraction methods include principal components regression and partial least squares regression.

3. Regularise the model by shrinking the highly parametrized model toward a less parametrized model. The aim of this approach, which includes Lasso and regularised discriminant analysis, is to lower the overall mean square error at the expense of increased bias. Such regularised methods are also referred to as "penalised discriminant analysis".

4. Finally, one can discover the $\hat{\mathbf{w}}$ for which $\|\hat{\mathbf{w}}\| = \sqrt{\hat{\mathbf{w}}\hat{\mathbf{w}}^t}$ is minimum for some least squares error threshold [Hand, 1997]. This approach is referred to as the shortest least squares solution and is distinct from regularisation which minimises a linear combination of the features plus a roughness penalty whose magnitude is adjustable. Solutions to the shortest least squares estimate include $\hat{\mathbf{w}} = (\mathbf{X^t X})^+ \mathbf{X^t y}$, where $(\mathbf{X^t X})^+$ is the Moore-Penrose generalised inverse of $\mathbf{X^t X}$ and algorithm based approaches such as the dynamic programing technique described by Kalaba et al. [1995].

Each of these methodologies is reviewed in turn.

4.3 Feature Selection and Subset Methods

The essence of subset selection according to Hastie et al. [2001] is to:

1. Improve prediction generalisability by setting some feature parameters to zero, and

2. Improve the interpretability of solutions by lowering the number of solution features.

Further, even when highly complex solutions are more accurate, one may be able to show that a smaller subset of features satisfactorily discriminates the classes. In applied terms one might suggest that in this case a limited number of *key* features are essentially "doing the work" of class discrimination with perhaps additional features included to either chase particular observations or to provide a small incremental improvement to the error estimate.

4.3.1 Best subset regression

This method finds the subset k features that yields the lowest residual sum of squares and is efficiently accomplished by the *leaps and bounds* algorithm of Furnival and Wilson [1974]. This algorithm can also be used to return the m best regressions rather than the single best solution.

4.4 Feature Extraction

Another approach to reducing the effective number of features is to extract or derive *de novo* features by transforming the original p-dimensional data to a lower dimensional subspace of $\mathbb{R}^p$. There are obviously many possible methods to extract such new features and one method, principal components regression is described here.

4.4.1 Principal Component Regression

Principal components regression seeks to map the original $\mathbb{R}^p$ data to a subspace which retains the underlying multivariate structure of the training set, $\mathcal{T}$. Assuming $\mathcal{T}$ is given by a mean-centred $\mathbf{X}$, the principal components (or Karhunen-Loeve directions) are given by the eigenvectors of the eigendecomposition (or spectral decomposition) of $\mathbf{X^t X}$ are given by the columns of $\mathbf{V}$ from 4.2 (p. pagerefSVD) [Golub and Van Loan, 1996].

The first principal component, $\mathbf{v_1}$, is the p-length vector in the direction of highest variance across data and each subsequent vector is orthogonal to all others in decreasing order of variance. Thus, the data projection $\mathbf{z_1} = \mathbf{X v_1}$ is the first principal component and has the highest variance of all linear combinations of the columns of $\mathbf{X}$ and

$$\text{var}(\mathbf{z_1}) = \text{var}(\mathbf{X v_1}) = \frac{\lambda_\mathbf{1}^\mathbf{2}}{N}$$

decreases as i increases such that $\mathbf{z}_{i=p}$ has the lowest variance.

By regressing the target value $\mathbf{y}$ onto $\mathbf{z_1}, \mathbf{z_2}, \ldots \mathbf{z}_m$ where $m < p$, we can derive the principal component regression parameters [Hastie et al., 2001],

$$\hat{\mathbf{w}}^{\mathbf{pcr}}(m) = \textstyle\sum_{i=\mathbf{1}}^{m} \hat{\theta}_m \mathbf{v}_m, \text{ where}$$
$$\hat{\theta}_m = \frac{\langle \mathbf{z}_m, \mathbf{y} \rangle}{\langle \mathbf{z}_m, \mathbf{z}_m \rangle}.$$

A comparison of the regression estimates between regularised methods (discussed below) and principal components regression is useful. Whereas the regularization shrinks $\hat{w}_i$ depending on the magnitude of the corresponding eigenvalue, principal components regression simply drops parameters for the $p - m$ smallest eigenvalues [Hastie et al., 2001]. However, as with ridge regression and other regularised techniques, principal components regression still uses all p features as even the first principal component is a p-length vector. Nevertheless, in practice, one may explore reduced subspaces of the data by setting low eigenvalues, e.g

$\lambda_i^2 \leq C \rightarrow 0$, and evaluating only the remaining parameters. Further, by comparing the eigenvalues between the first and subsequent principal components, one can explore relative feature "importance" in the extracted feature space.

4.5 Regularization and Penalization Methods

Also known as *shrinkage* methods, these solutions attempt to control model parameter complexity by placing a penalty on the size or number of parameters. The degree of shrinkage is controlled by a penalty parameter. We start by considering ridge regression technique developed by Hoerl [1962] and later described in Hoerl and Kennard [1968, 1970]. For comparison, we will also discuss the Lasso regularisation introduced by Tibshirani [1996].

4.5.1 Ridge regression

Ridge regression was developed by Hoerl with the aim of controlling the magnitude of parameter estimates $\hat{\mathbf{w}}_i$ and improving stability of the ordinary least squares solution as $\mathbf{X}^{\mathrm{t}}\mathbf{X}$ becomes more ill-conditioned [Hoerl and Kennard, 1970]. In the ridge regression formulation, model complexity is controlled by imposing a penalty on the ordinary least squares solution by penalising the length of $\hat{\mathbf{w}}$ weighted by λ. The penalty is realised by adjusting the cross-product matrix $\mathbf{X}^{\mathrm{t}}\mathbf{X}$ used in the estimate $\hat{\mathbf{w}}$ by the addition of $\lambda\mathbf{I}$, where $\mathbf{I}$ is the identity, such that

$$\hat{\mathbf{w}}^{ridge} = [\mathbf{X}^{\mathrm{t}}\mathbf{X} + \lambda\mathbf{I}]^{-1}\mathbf{X}^{\mathbf{t}}\mathbf{y};\ \lambda \geq 0.$$

Hoerl and Kennard [1970] show that the ordinary least squares estimate $\hat{\mathbf{w}}^{ols}$ can be related to the ridge solutions as

$$\hat{\mathbf{w}}^{\mathbf{ridge}} = [\mathbf{I}_p + \lambda(\mathbf{X}^{\mathbf{t}}\mathbf{X})^{-\mathbf{1}}]^{-1}\hat{\mathbf{w}}^{ols}.$$

One manner in which this penalty imposes stability (regularises) and lowers the expected variance of the solution is by controlling the degree to which a very large positive parameter value can offset a very large negative parameter value [Hastie et al., 2001]. Also, by adding a positive constant λ to the diagonal of $\mathbf{X}^{\mathrm{t}}\mathbf{X}$ the solution is guaranteed to be non-singular and thus invertible [Nakos and Joyner, 1998].

As with the LDA solution presented in the previous chapter, the ridge solution benefits from factoring $\mathbf{A}$=$\mathbf{X}^{\mathrm{t}}\mathbf{X}$, using a singular value decomposition

$$\mathbf{A}= \mathbf{U\Lambda V^{t}}.$$

For comparison we first use this decomposition to derive the ordinary least squares solution which is unbiased and is the minimum "Gauss-Markov" linear solution:

$$\begin{aligned}
\mathbf{X\hat{w}^{ols}} =&\mathbf{X(X^{t}X)^{-1}X^{t}y}\\
=&\mathbf{U\Lambda V^{t}[(U\Lambda V^{t})^{t}(U\Lambda V^{t})]^{-1}(U\Lambda V^{t})^{t}y}\\
=&\mathbf{U\Lambda V[(V\Lambda U^{t})(U\Lambda V^{t})]^{-1}(V\Lambda U^{t})y}\\
=&\mathbf{U\Lambda V^{t}[V\Lambda\Lambda V^{t}]^{-1}V\Lambda U^{t}y}\\
=&\mathbf{U\Lambda V^{t}(V^{t})^{-1}\Lambda^{-1}\Lambda^{-1}V^{-1}V\Lambda U^{t}y}
\end{aligned}$$

and, using the above,

$$\begin{aligned}
=&\mathbf{U\Lambda\Lambda^{-1}\Lambda^{-1}\Lambda U^{t}y}\\
=&\mathbf{UU^{t}y}.
\end{aligned}$$

For ridge regression,

$$\begin{aligned}
\mathbf{X\hat{w}}^{ridge} =&\mathbf{X(X^{t}X+\lambda I)^{-1}X^{t}y}\\
=&\mathbf{U\Lambda V^{t}[(U\Lambda V^{t})(U\Lambda V^{t})+\lambda I]^{-1}(U\Lambda V^{t})^{t}y}\\
=&\mathbf{U\Lambda V^{t}[(V\Lambda U^{t})(U\Lambda V^{t})+\lambda I]^{-1}V\Lambda U^{t}y}
\end{aligned}$$

$$
\begin{aligned}
&= \mathbf{U\Lambda V^t}[(\mathbf{V\Lambda\Lambda V^t}) + \lambda\mathbf{I}]^{-\mathbf{1}}\mathbf{V\Lambda U^t y} \\
&= \mathbf{U\Lambda V^t}[(\mathbf{V\Lambda^2 V}) + \lambda\mathbf{V V^t}]^{-\mathbf{1}}\mathbf{V\Lambda U^t y} \\
&= \mathbf{U\Lambda V^t}[\mathbf{V}(\mathbf{\Lambda^2} + \lambda\mathbf{I})\mathbf{V^t}]^{-\mathbf{1}}\mathbf{V\Lambda U^t y} \\
&= \mathbf{U\Lambda V^t}(\mathbf{V^t})^{-\mathbf{1}}(\mathbf{\Lambda^2} + \lambda\mathbf{I})^{-\mathbf{1}}\mathbf{V}^{-\mathbf{1}}\mathbf{V\Lambda U^t y} \\
&= \mathbf{U\Lambda}(\mathbf{\Lambda^2} + \lambda\mathbf{I})^{-\mathbf{1}}\mathbf{\Lambda U^t y} \\
&= \frac{\mathbf{U\Lambda^2 U^t y}}{\mathbf{\Lambda^2} + \lambda\mathbf{I}} \\
&= \sum_{i=1}^{N} \mathbf{u}_i \frac{d_i^2}{d_i^2 + \lambda} (\mathbf{u}_i)^{\mathbf{t}}\mathbf{y}.
\end{aligned}
$$

This derivation illustrates that as the singular values d_i^2 become smaller for a given basis vector $\mathbf{u}_i$ the shrinkage of $\hat{w}_i^{ridge}$ will increase. Consequently, directions in the column space of $\mathbf{X}$ with the smallest variance will be most affected by the regularization. This approach makes an implicit assumption that the response of interest (e.g. phenotype) will vary the most *between classes* in the direction of the highest variance. While this may be a generally reasonable assumption, one could explore this relationship for a particular data set and, at the very least, we suggest that this assumption should be explicitly considered if ridge regression is applied. In practice, one might, for example, view the data projected into the first principal components highlighted by class to satisfy ourselves that the ridge regularization is appropriate in a particular case.

An alternative presentation is useful for comparing ridge regression to other penalized regression methods. Whereas the residual sum of squares (RSS) estimate $\hat{\mathbf{w}}$ is found by minimising

$$
\min(RSS(\mathbf{w})) = \underset{\mathbf{w}}{\operatorname{argmin}} \left[\sum_{i=1}^{N} (y_i - w_0 - \sum_{j=1}^{p} x_{ij} w_j)^2 \right],
$$

the ridge regression solution adds the additional penalty term of the squared parameter

$$
\hat{\mathbf{w}}^{ridge} = \underset{\mathbf{w}}{\operatorname{argmin}} \left[\sum_{i=1}^{N} (y_i - w_0 - \sum_{j=1}^{p} x_{ij} w_j)^2 \right] + \lambda \sum_{j=1}^{p} w_j^2.
$$

Finally, for $\lambda = 0$, the ridge solution reduces to the ordinary least squares result.

4.5.2 The Lasso

The lasso regression was designed to address the deficiencies of both ridge regression and subset regression. While ridge regression may improve the mean squared error estimate of an ordinary least squares solution by shrinking coefficients, the results do not improve interpretability as p is essentially unchanged because no coefficients are reduced to zero (i.e. removed). Subset regression, on the other hand, is more interpretable for discrete $k < p$ but can be relatively unstable as small perturbations in the data $\mathcal{T}$ can lead to very different models [Tibshirani, 1996].

The method introduced by Tibshirani [1996] builds from, and improves, the non-negative garrote introduced by Breiman [1993] by minimizing the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant t which controls shrinkage:

$$\hat{\mathbf{w}}^{lasso} = \text{argmin} \left[\sum_{i=1}^{N} (y_i - w_0 - \sum_{j=1}^{p} x_{ij} w_j)^2 \right],$$

$$\text{subject to } \sum_{j=1}^{p} |w_j| \leq t.$$

The lasso constraint is thus incorporated to the *RSS* solution in the Lagrangian form

$$\hat{\mathbf{w}}^{lasso} = \min_{\mathbf{w}} \left[\sum_{i=1}^{N} (y_i - w_0 - \sum_{j=1}^{p} x_{ij} w_j)^2 \right] + \lambda \sum_{j=1}^{p} |w_j| .$$

Thus, if λ is chosen sufficiently large, some of the parameters will reduce to zero [Tibshirani, 1996, Bishop, 2006]. This method provides the benefits of both subset selection and ridge regression without the associated disadvantages

[Tibshirani, 1996]. The method enjoys both the stability of regularised regression and is interpretable as the effective p is reduced as coefficients shrink to zero.

To solve the lasso a quadratic programming algorithm is required and an efficient algorithm was introduced by Tibshirani [1996].

4.6 Shortest Least Squares

We conclude this review of approaches to dealing with the $p > n$ case with a description of the shortest least squares solution. In the case where $\mathbf{X^t X}$ is of reduced rank, the standard equation

$$\hat{\mathbf{w}} = (\mathbf{X}^{\mathrm{t}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{t}}\mathbf{y},$$

has no solution. On the other hand, we may force a unique solution by stipulating additional constraints. The shortest least squares method finds that unique solution $\hat{\mathbf{w}}$ that minimises the total parameter length, $\|\hat{\mathbf{w}}\|$. This solution is found by replacing $(\mathbf{X^t X})^{-1}\mathbf{X^t}$ with the Moore-Penrose pseudo-inverse defined [Duda et al., 2000] in general form as

$$\mathbf{A}^{+} = \lim_{\lambda \to 0} (\mathbf{A^t A} + \lambda \mathbf{I})^{-1}\mathbf{A^t},$$

which yields the $\hat{\mathbf{w}}^{sls}$ estimate

$$\hat{\mathbf{w}}^{sls} = \mathbf{X}^{+}\mathbf{y}.$$

There are two interesting points we can make with regard to this solution:

1. If $\mathbf{X^t X}$ is of full rank where, e.g. $N = p = rank(\mathbf{X})$, then the pseudo-inverse is equal to the general inverse, $\mathbf{X}^{+} = \mathbf{X}^{-1}$, and the solution is $\hat{\mathbf{w}} = \mathbf{X}^{-1}\mathbf{y}$ [Golub and Van Loan, 1996].

2. We also note that the pseudo-inverse solution has a strong relationship to the ridge solution above and, in fact, for $\lambda = 0$ and in the case of full rank:

$$\hat{\mathbf{w}}^{ols} = \hat{\mathbf{w}}^{ridge} = \hat{\mathbf{w}}^{sls}.$$

4.7 Conclusions

In this chapter a number of commonly used techniques to handle the $p >> N$ case were introduced, i.e. where the number of variables greatly exceeds the number of observations. This case is applicable in the analysis of nearly all microarray data sets because the number of patient specimens is at most several hundred specimens while the number of genes or transcripts measured is typically over ten thousand.

In the microarray analysis carried out in this research, for example, the number of specimens is typically in the range of 1-4$\times 10^2$ while the number of probesets is in the range of 4-5$\times 10^4$. The methods introduced in this review are applied in this research with a particular emphasis on subset selection.

Chapter 5

Materials and Methods

5.1 Aims

This chapter describes the materials and methods used in the discovery and hypothesis testing of candidate gene expression biomarkers for colorectal neoplasia. Standard statistical techniques and bioinformatics tools are also discussed as well as extensions to these tools that were developed by the author.

All discovery data were acquired in collaboration with third parties. After carefully checking data using established and newly developed quality control metrics, a range of mathematical techniques were applied to these data in order to discover candidate biomarkers. Finally, each hypothetical biomarker candidate was tested using clinical specimens independently obtained for this project. The biological validation data reported here were generated in the lab either by the author or using contracted third-party assistance using standard protocols under guidance from the author.

5.2 Discovery data

5.2.1 Differential display discovery

The first set of candidate gene expression markers were discovered in collaboration with a team lead by Dr. Rob James and Prof. Graeme Young at the Flinders University from 1999-2001 using differential display PCR technology. To discover tumour associated RNA species, Dr. James' lab extracted total RNA from adenomatous polyps (including tubular, tubulovillous, and villous adenomas) and adenocarcinoma from colonoscopy or surgical specimens collected under a Flinders Ethics Committee Approved Protocol. Non-neoplastic controls confirmed to be normal by histopathology review were also collected. All tissues were snap-frozen on dry-ice and RNA purification columns were employed to isolate RNA.

Differential display PCR was carried out by Dr. James and his colleagues using the Hieroglyph mRNA profile kit (Genomyx Corp, Foster City, CA USA) which involved a first strand cDNA synthesis using a 12-set combination of 2-base anchor primers (containing a 17 nucleotide T7 promoter sequence) and 20 arbitrary upstream primers. PCR reactions were performed in the presence of ^{32}P labeled ATP. PCR products for RNA from both normal and disease specimens were separated on 61cm 4.5% polyacrylamide sequence gel under denaturing conditions. Differentially expressed bands that showed evidence of increased expression in adenomas were excised from the gel, eluted, re-amplified and sequenced. Candidate expression targets were cloned into a cDNA library and further screened and sub-cloned to isolate sequences of interest. Finally, full-length transcripts for candidate biomarkers from the study were determined by either cDNA library screening or Rapid Amplification of cDNA ENDS (5' or 3' RACE).

From a total of 1145 bands identified to be over-expressed in adenomas, a subset of 354 targets were chosen for initial quantitative reverse transcriptase PCR screening in a pilot scale study using pooled samples. A pilot study measuring these 354 transcripts by RT-PCR in 15 pooled neoplastic tissues (5 tubular

adenomas, 5 tubulovillous adenomas, and 5 adenocarcinomas) and five (5) non-neoplastic controls confirmed that 67 (19%) transcripts showed 10-fold or higher expression level in the neoplasias.

The subset of 67 targets exhibiting 10-fold or greater up-regulation was measured by quantitative RT-PCR in an expanded set of clinical specimens. This study measured the expression profile of each of the 67 candidates using specifically designed primers in independent test tissues including 51 adenomas (21 tubular adenomas, 26 tubulovillous adenomas, 4 villous adenomas) and 20 non-neoplastic controls. Adenocarcinoma specimens were not included in this experiment to maximise the number of adenomas under consideration. For this experiment, the amplification reaction cycles were measured by real-time fluorescence until a threshold fluorescence intensity is measured. Thus, the higher the initial concentration of RNA, the lower the number of cycles required to reach a given threshold. These RT-PCR results (thresholded cycle values) were normalised using beta-actin *ACTB* transcript expression and provided to the author for analysis.

After removing redundant targets, 328 of the 354 RNA targets formed the basis of an international patent submission in 2001 (on which the author is a co-inventor) that established claims against these 328 candidate biomarkers [James, 2001]. In particular, the patent claims are based around the sequence data combined with the author's multivariate analyses which prioritise and demonstrate clinical utility of the candidate sequence.

5.2.2 GeneLogic data

Gene expression profiling data measured in 548 colorectal tissue specimens and accompanying clinical data was purchased from GeneLogic Inc (Gaithersburg, MD USA). Raw oligonucleotide microarray data totalling 44,928 probesets (Affymetrix HGU133A & HGU133B, combined), experimental and clinical descriptors, and digitally archived microscopy images of histological preparations were received for each of the 548 tissues.

A full list of covariate data provided for each tissue is shown in Appended table D.1.

These data were generated using oligonucleotide microarrays hybridised to labeled cRNA synthesised from poly-A mRNA transcripts isolated from colorectal tissue specimens. These microarrays (GeneChips) were processed by GeneLogic according to the manufacturer's (Affymetrix Inc) instructions with particular attention paid to reproducible, industrial standard lab protocols.

Prior to carrying out discovery research using these data, extensive quality control testing was applied. The quality control methods, some of which are novel, are described in detail in Appendix B. In addition to the quantitative assessment of expression data, each data record was manually assessed for clinical consistency and a sample of tissue specimens was randomly chosen for histopathology audit by a clinical expert using the digital histology images provided.

After all quality control routines were applied, a total of 454 microarrays were judged suitable for our research purposes. A phenotypic breakdown of these chips is shown in 5.1

Table 5.1: An analysis of specimen phenotypes comprising the discovery microarray data purchased from GeneLogic judged suitable for research purposes after extensive quality control testing.

		Normal	IBD	Adenoma	Cancer
Gender	Female	102	17	16	93
	Male	120	25	13	68
Anatomy	Proximal	70	13	13	58
	Distal	95	12	5	90
	Unknown	57	17	11	13
Age	under 40	26	15	3	8
	40-49	22	13	3	21
	50-59	39	8	5	34
	60-69	53	5	8	41
	70-79	52	1	7	34
	over 79	30	1	3	23

5.3 Validation data:

5.3.1 Custom microarray

To test the many hypothetical gene biomarkers identified by discovery research and analyses, a custom-designed microarray was employed. After a cursory review of alternative microarray-based technologies, the Affymetrix oligonucleotide microarray platform was chosen. Key considerations of this decision included 1) the availability of microarray testing equipment and hardware for use by the author; 2) the availability of numerous publicly-available analytical data mining tools provided for the R statistics language; and 3) a growing acceptance of Affymetrix-based microarrays in both the scientific literature as well as for commercial research.

The oligonucleotide content of the custom chips was based on three discovery sources:

1. Transcript nucleotide sequences discovered by differential expression analysis and in-house sequencing experiments conducted in 1999.
2. Probesets and gene symbols discovered using a commercially available data set purchased from GeneLogic Inc.(USA) that includes 454 samples interrogated by full-genome GeneChips.
3. Candidate markers described in the literature.

Following discussions with Affymetrix, the "Custom AffyExpress" microarray program was chosen to construct the custom design oligonucleotide microarray for hypothesis testing based on cost and suitability of design.

5.3.2 Microarray geometry and design considerations

The Affymetrix oligonucleotide microarray (GeneChip®) is manufactured by photolithographic technologies analogous to those used to create silicon based integrated circuits and microchips. Table 5.2 provides an overview of design

options for the custom AffyExpress microarray design. From these options the #100-3660 format was used for fabrication. Using the 11μm design, this format provided up to 4,800 targets, assuming the standard 22 probes per probeset or 9,600 perfect match-only probesets.

Table 5.2: Geometry options for design.

Feature size	11 microns	
Probesets per chip	1,700	(#100-2187 format)
	4,800	(#100-3660)
	10,000	(#49-5241)
	23,000	(#49-7875)
Probes per set	22 (nominal)	
Bases per probe	25 (nominal)	
Total features	105,600	(based on a #100-3660 format)

5.3.3 Perfect match (PM) vs. mismatch (MM) probes

The standard commercial microarray manufactured by Affymetrix specifies 22 probes per target. These 22 probes include 11 "perfect" match 25-mer oligonucleotide probes designed to hybridise with a 25-mer span of the labelled cRNA target. The remaining 11 probes are identical to the perfect match probes except for the middle (13th) base, which is Watson-Crick reversed. These "mismatch" probes are generally included to provide, in theory, a means to estimate the cross-hybridisation of non-target cRNA to a given probe. The entire set of 22 probes is then randomly scattered across the microarray surface to avoid regional bias introduced during the hybridisation.

Based on a review of background and normalization methods used for microarray experiments, the utility of mismatch probes is questionable [Irizarry et al., 2003]. Further, statistical models for multiple array binding data using only the 11 perfect match probes and normalised using the "robust multichip array" algorithm developed by Irizarry et al. provide consistently better observed to expected performance based on artificial (controlled) spike-in experiments [Irizarry et al., 2003].

Consequently, the microarray fabrication specification used for these validation experiments did not include mismatch probe content. By removing these probes, the custom validation microarray provides space for up to 9,600 probesets of 11 probes per set, or twice the "usual" content of 4,800 probesets of 22 probes per set.

5.3.4 Labelled cRNA vs. cDNA

A key consideration in the design of the validation microarray was to employ a technology suitable for measuring specific transcript exons which are not biased to the 3' mRNA terminus. In particular, the differential display markers were discovered using a randomly primed transcriptome unlike the GeneLogic microarray data which consisted of probesets designed to hybridise to the terminal 600 bases at the 3' end of RefSeq-based GenBank sequences.

This consideration concerns the choice of amplification and labelling protocol used to synthesise the labelled product for hybridisation to the microarray. The conventional protocols and kits supplied with Affymetrix commercial arrays, e.g. the HG U133plus2 (whole genome) array, use a T7 Oligo(dT) promoter to prime the first reverse transcription of total RNA isolated from a sample. After a second strand cDNA synthesis, the dsDNA is *in vitro* transcribed in the presence of biotinylated bases to create a labelled cRNA product that is hybridised to the microarray and measured. This biotin-labelled cRNA is the ANTISENSE strand of the target.

The printed cDNA probesets, therefore represent the SENSE strand of the target. As the differential display targets were discovered by amplifying randomly primed adenoma transcriptome targets, many of these targets do not have a poly-A tail. The usual strategy of using the poly-dT labelling procedure is therefore not suitable.

Affymetrix recently introduced an alternative labelling protocol for their commercial "Exon Array" products (i.e. whole transcriptome microarrays) that

makes use of a T7-N6-primer of random hexamers. The requirement for labelling protocols that do not mandate a well formed poly-A tail message suggests that the new random hexamer-based protocol that prints ANTISENSE strand cDNA on the microarray is preferred. This methodology provides continuity with both the oligonucleotide discovery data and the differential display discovery data. For specific details of the labelling methods used for the custom microarrays see 5.4.3.

5.4 Laboratory methods

An overview of methodologies used in the laboratory to carry out this research is presented.

5.4.1 Human tissue samples

For all hypothesis testing experiments, independently collected clinical samples were obtained from a tertiary referral hospital tissue bank in metropolitan Adelaide, Australia (Repatriation General Hospital and Flinders Medical Centre). Access to the Tissue Bank for this research was approved by the Research and Ethics Committee of the Repatriation General Hospital and the Ethics Committee of Flinders Medical Centre. Informed patient consent was received for each tissue studied.

Following surgical resection, specimens were placed in a sterile receptacle and collected from theatre. The time from operative resection to collection from theatre was variable but not more than 30 minutes. Samples, approximately 125mm3 (5x5x5mm) in size, were taken from the macroscopically normal tissue as far from neoplastic pathology as possible, defined both by colonic region as well as by distance either proximally or distally to the pathology. Tissues were placed in cryovials, then immediately immersed in liquid nitrogen and stored at -150°C until processing. Clinical data were available for each specimen examined,

including histopathological diagnoses related to the specimens tested and the site in the colorectum from which the material was derived.

5.4.2 RNA extraction

Frozen samples were processed either by the author using (Method I) or under commercial contract by Flinders Medical Centre staff (Method II) using standard protocols and commercially available kits. Each fresh frozen specimen was carefully dissected to maximise the epithelial portion of extracted portion. No attempt was made, however, to micro-dissect epithelial tissue exclusively as molecular markers might derive from non-epithelial (e.g. stromal) tissue as well as epithelial tissue.

Method I

Method I was used to extract RNA for use from the specimens used for testing transcripts differentially expressed along the longitudinal axis of the colon described in 6.

Briefly, frozen tissues were homogenised using a carbide bead mill (Mixer Mill MM 300, Qiagen, Melbourne, Australia) in the presence of chilled Promega SV RNA Lysis Buffer (Promega, Sydney, Australia) to neutralise RNase activity. Homogenised tissue lysates for each tissue were aliquoted to convenient volumes and stored -80°C . Total RNA was extracted from tissue lysates using the Promega SV Total RNA system according to manufacturer's instructions and integrity was assessed visually by gel electrophoresis.

Method II

Method II was used to extract RNA for microarray experiments to test hypotheses related to biomarker candidates for colorectal neoplasia described in Chapter 8.

RNA extractions were performed using Trizol®reagent (Invitrogen, Carlsbad, CA, USA) as per manufacturer's instructions. Each sample was homogenised in 300μL of Trizol reagent using a modified Dremel drill and sterilised disposable pestles. Additional 200μL of Trizol reagent was added to the homogenate and samples were incubated at room temperature (25°C) (RT) for 10 minutes. 100μL of chloroform was then added, samples were vortexed for 15 seconds, and incubated at RT for 3 further minutes. The aqueous phase containing target RNA was obtained by centrifugation at 12,000 rpm for 15 min, 40°C . RNA was then precipitated by incubating samples at RT for 10 min with 250μL of isopropanol. Purified RNA precipitate was collected by centrifugation at 12,000 rpm for 10 minutes, 40°C and supernatants were discarded. Pellets were then washed with 1ml 75% ethanol, followed by vortexing and centrifugation at 7,500g for 8 min, 40°C . Finally, pellets were air-dried for 5 min and re-suspended in 80μL of RNase free water. To improve subsequent solubility samples were incubated at 55°C for 10 min. RNA was quantified by measuring the optical density at A260/280 nm. RNA quality was assessed by electrophoresis on a 1.2% agarose formaldehyde gel.

5.4.3 Microarray processing

HG U133 Plus 2.0 GeneChips

To measure relative expression of mRNA transcripts from along the longitudinal axis of the colon, RNA extracts from non-neoplastic tissue with known site of origin along the large intestine were analyzed using Affymetrix HG U133 Plus 2.0 GeneChips (Affymetrix, Santa Clara, CA USA) according to the manufacturer's protocols. Briefly, biotin-labeled cRNA was prepared using 5μ (1.0 μg/μL) total RNA (approx. 1 μg mRNA) with the "One-Cycle cDNA" kit (incorporating a T7-oligo(dT) primer) and the GeneChip IVT labeling kit. In vitro transcribed cRNA was fragmented (20 μg) and analyzed for quality control purposes by spectrophotometry and gel electrophoresis prior to hybridisation. Finally, an hybridisation cocktail was prepared with 15μg of cRNA (0.5 μg/μL) and hybridised

to HG U133 Plus 2.0 microarrays for 16h at 45°C in an Affymetrix Hybridisation Chamber 640. Each cRNA sample was spiked with standard prokaryotic hybridisation controls for quality monitoring.

CG_AGP custom microarray

To test hypotheses related to biomarker candidates for colorectal neoplasia RNA extracts were assayed using a proprietary gene chip designed by the author in collaboration with Affymetrix (model designation: CG_AGPa520460F) and further described in section 5.3. These assays were processed by CSIRO technicians (North Ryde, NSW) under commercial contract. Importantly methods were initially developed by, and all work was supervised by, the author.

These validation microarrays were processed using the standard Affymetrix protocol developed for the HuGene ST 1.0 array described in [Affymetrix, 2007]. This method was chosen because target RNA is randomly primed using random hexamer primers instead of the poly-dT method (described above) which was used for both the GeneLogic discovery data and the HG U133plus2 microarrays. The selection of a randomly priming methodology was important for two reasons. First, only a random priming of the transcriptome is suitable for the differential display discovery biomarkers which were not discovered by 3' (i.e. poly-A) techniques. The use of a non-3' biased labeling method also provides a further layer of technical validation for the probesets discovered by "standard" poly-dT protocols used for the HG U133A & B probesets from the GeneLogic discovery data.

According to the HuGene ST 1.0 protocol first cycle, dsDNA was synthesised from 100ng of total RNA extract using random hexamer primers tagged with T7 promoter sequence and SuperScript II (Invitrogen, Carlsbad CA) and then DNA Polymerase I. Anti-sense cRNA was then synthesised using T7 polymerase and combined with SuperScript II, dUTP (+dNTP), and random hexamers to synthesise sense strand cDNA incorporating uracil. A combination of uracil DNA glycosylase (UDG) and apurinic/apyrimidinic endonuclease1 (APE 1) were

used to fragment the DNA product. Next, the DNA was biotin labelled by terminal deoxynucleotidyl transferase (TdT) with the Affymetrix proprietary DNA Labeling Reagent covalently linked to biotin. Hybridisation to the Custom microarray CG_AGPa520460F was carried out at 45°C for 16-18h. Finally, the microarrays were washed, stained and scanned as above.

All microarrays were stained with streptavidin phycoerytherin and washed with a solution containing biotinylated anti-streptavidin antibodies using the Affymetrix Fluidics Station 450. Finally, the stained and washed microarrays were scanned with the Affymetrix Scanner 3000.

5.4.4 RT-PCR

Quantitative real time polymerase chain reaction (RT-PCR) was used to confirm selected gene expression discoveries using Applied Biosystems pre-designed and optimized TaqMan gene expression assays. These RT-PCR data were collected by Dr. Glenn Brown, CSIRO Molecular Health Technologies (North Ryde, NSW), however experimental design including tissue and target gene selection was carried out by the author. The selection of particular tissue specimens were balanced for gender, age, and proximal-distal origin as appropriate.

Prior to RT-PCR analysis, 100ng of total RNA was subject to linear amplification using the QIAGEN QuantiTect Whole Transcriptome amplification kit (QIAGEN, USA) according to the manufacturer's instructions. 2.0μl of the amplified, diluted (1:50) cDNA was then analysed in a 25μl reaction volume by RT-PCR using TaqMan universal master mix (Applied Biosystems, USA) in an ABI prism 7700 sequence detector (Manufacturer, Country) following manufacturer's protocols. These assays were performed in triplicate and resulting expression levels were quantified as a ratio to three "housekeeping" genes (*HPRT*, *TBP* and *GAPDH*). These genes are often used in colorectal RT-PCR experiments because of their relatively low variance in expression levels measured in most colorectal tissue phenotypes. Final quantified results were reported using the Δ-cycle threshold method.

5.5 Statistical methods

5.5.1 Statistical software and data processing

The R statistics environment was used for most statistical analyses [R Development Core Team, 2008] and open source libraries from BioConductor (BioConductor, www.bioconductor.org) [Gentleman et al., 2004] were used for analysing microarray data. Custom software was written in a range of languages and tools. C++ was used for implementing the support vector machines algorithm and for all-subsets analysis of differential display data using the k-nearest neighbor metric. Perl and C was used for databasing and automated sequence annotation, in particular. Bioinformatics tools written in Perl often utilised open source libraries provided by BioPerl [Stajich et al., 2002]. All data processing was performed on Unix desktop variants including MacOSX and Unix.

5.5.2 Affymetrix GeneChip data reduction

The Affymetrix GCOS software package was used to transform raw microarray image files created by the Affymetrix Scanner to a digitized format. Raw CELDATA files were processed using either manufacture or custom chip description files (CDFs) as appropriate. CDFs for GeneLogic HG U133A&B discovery data and the HG U133plus2 data were downloaded from BioConductor and Affymetrix, respectively. The CDF used for processing the custom validation microarray was created by the author using open source libraries for R and manufacturing probeset coordinate files provided by the design team at Affymetrix.

Gene expression levels were calculated by both Microarray Suite (MAS) 5.0 (Affymetrix) and the Robust Multichip Average (RMA) normalization techniques [Affymetrix, 2004a, Irizarry et al., 2003]. MAS normalised data was used for accessing standard quality control routines only.

All discovery and hypothesis testing data were normalised using the RMA algorithm implemented in R. This algorithm involves three discrete steps. First

the raw data are background corrected for both optical noise and non-specific oligonucleotide binding. Next the data are transformed onto a log base-2 scale and each probe (not probeset) is quantile normalised across all microarrays. Finally probesets are constructed by aggregating quantile normalised probes using median polish [Irizarry et al., 2003].

5.5.3 Annotation of discovery data

BLAST-based annotation of differential display sequences

The results of differential display discovery detailed in 7 produced a set of 328 nucleotide sequences encoding hypothetical RNA biomarker candidates with observed higher expression in colorectal adenomas relative to normal control tissues. Since elucidating these genomic sequences in 2001, annotation of these sequences to putative gene target has based on GenBank sequence data has evolved considerably.

To automate the routine alignment of these proprietary candidate nucleotide sequences with the expanding GenBank database, automated BLAST and parsing tools were written in Perl and C by the author. Briefly, these automated tools submit each sequence to NCBI for sequence alignment using nBLAST. nBLAST reports were then parsed and GenBank hits were "graded" according to a subjective ranking which prioritised alignment "hits" according to percent alignment, coverage, species, etc. The parser also graded hits by searching for keywords in the GenBank record description including e.g. 'refseq', 'hypothetical gene'. Finally, each GenBank record was translated to one or more official gene symbols via a look-up table downloaded from GenBank and all results were stored into a MySQL database developed by the author. A human readable report was produced for manual human review of the final auto-generated results.

These annotation results were critical to a) the design of the custom gene chip in terms of accessing commercially available probesets targeting known gene transcripts; and b) to the ability to compare discovery (and testing) results

between the differential display research and the Affymetrix probeset data.

HG U133 (A/B/Plus2) annotation

Affymetrix oligonucleotide microarray data were analysed at the 'probeset' level. One should note, however, that annotation maps between Affymetrix probeset ID and putative gene symbols are not static. While the HG U133 data sets and the custom microarray were engineered based on the latest available Unigene cluster data for human gene transcript and exon data *at the time of product design*, knowledge of the human genome is in a constant state of flux. As a consequence, probeset ID to gene symbol mappings were observed to change for some probesets from time to time. This dynamic binding is a general challenge of working with microarray data and the issue is not particular to this research. While the latest available metadata mapping available from BioConductor (and Affymetrix) were used at the time of each *individual* analysis, there is a possibility for minor inconsistencies in reporting over the course of this research which describes analyses carried out over approximately three years. Nevertheless, the latest available annotation was used for the design of the custom validation microarray which was perhaps the most time sensitive component of this research. The most recent annotations are also used for this thesis.

An additional consideration is that individual 25-mer Affymetrix probes which are aggregated to probeset level reporting do not hybridise to target cRNA transcripts to the same degree of specificity. While Affymetrix makes a strong effort to choose sequence specific oligonucleotide probes, the homologous sequences within some genes introduces a degree of promiscuity. As a consequence, this work biases biomarker selection and reporting toward higher specificity probesets where possible and promiscuous probesets are identified when appropriate. The Affymetrix probeset naming conventions are useful in understanding cross-hybridisation potential, shown in Table 5.3 [Affymetrix, 2004a]:

Table 5.3: Probeset naming conventions

Naming Style	Interpretation
probeset_at	Probeset is unique for a target transcript.
probeset_s_at	Probes may be shared by two or more transcripts. In most cases these probesets target multiple transcripts from the same gene (e.g. splice variants), but they can also potentially bind to homologous genes. The probesets are all common to the multiple transcripts.
probeset_x_at	Probeset contains probes that are identical to or similar to unrelated transcript sequences. These probesets may bind in an unpredictable manner.

Custom microarray annotation

The `hgu133plus2` library version 2.2.0 was used to map probeset IDs to gene symbol on the custom validation microarray for this thesis. This library was assembled using Entrez Gene data downloaded on Apr 18 12:30:55 2008 [Gentleman et al., 2004].

Earlier aspects of this research, e.g. related to the gene expression map along the large intestine, were annotated using `hgu133plus2` library version 1.16.0 (15 March 19:46 2007), or earlier.

5.5.4 Hypothesis testing of differentially expressed biomarkers

To assess differential expression between tissue classes, Student's t test for equal means between two samples as implemented in the "limma" library [Smyth, 2004][Smyth, 2005] was used. To mitigate the impact of false discovery due to multiple hypothesis testing, significance levels (P values) were adjusted according to Bonferroni in the discovery process [Bland and Altman, 1995]. The Benjamini & Hochberg correction for controlling the false discovery rate of solutions[Benjamini and Hochberg, 1995] was used for analyses in the validation data set.

5.5.5 Inter-segment modeling of the large intestine

To evaluate the nature of inter-segment gene expression along the colorectum, probesets that were differentially expressed between the terminal segments (i.e. caecum vs. rectum) were analyzed for relative fit to linear models in a multi-segment (i.e. caecum, ascending, descending, sigmoid and rectum) versus a two segment framework. The goal of this analysis was to explore whether such probesets are better modelled by a five-segment linear model that approximates a continual gradation or by a simpler, dichotomous "proximal" vs. "distal" gradient.

As these data were only identified by colorectal segment designation and not by a continuous measurement along the length of the colon, the continuous model could only be approximated using the tissue segment location. Probesets that were differentially expressed between the most terminal segments (caecum and rectum) were used for this analysis in order to maximize the likelihood of identifying transcripts that varied along the proximal-distal axis of the colon.

Probeset expression levels were first modelled along the proximal-distal axis of the colon using a five factor linear model according to an indicator matrix defined by the colorectal segment for each tissue. For this model each tissue was assigned by removal location exclusively to one of: caecum, ascending, descending, sigmoid, or rectum. Transverse tissues were not included because such tissues could not be *a priori* assigned to either the distal or proximal region. This difficulty arises because intra-segment locations for tissues were not provided and because the hypothetical divide of the proximal-distal is approximately two thirds the length of the transverse segment.

This five segment model was then compared to a two-factor robust linear model corresponding to the theoretical proximal and distal regions of the colon. Thus, for the two segment model, the first factor (corresponding to the proximal tissues) included all of the tissues from the caecum and ascending colon while the second factor (corresponding to the distal colon) included all tissues from the descending, sigmoid and rectum segments.

When comparing these distinct models for each probeset, an F-test was used to evaluate the alternative hypothesis that the improved fit (reduced regression residual) provided by the more complex five-segment model was significantly better than the simpler two segment model. A non-significant residual reduction indicated a failure to reject the null-hypothesis so that there would be no inherent value in adopting a more complex five segment model over the simpler alternative.

5.5.6 Logistic regression modeling

Except in rare circumstances (e.g. multi-segment modeling discussed above), this research tested two-class comparisons exclusively. For example, analyses were carried out for normal tissues vs. adenomas, or normal vs. cancer, or neoplastic vs. non-neoplastic phenotypes. These two-state discriminants (i.e. the predicted is either class "A" or class "B", exclusively) were conveniently modelled using a regression model that restricts the response to a [0,1] range. The logistic regression is most widely used to satisfy this criterion [Hand, 1997] and is modelled as a linear combination of the logistic transform of the class probability, $\log(P(\mathbf{C_1})/(1 + P(\mathbf{C_1})))$.

Logistic regression models were routinely used to assess and compare classification models involving one or more biomarker variables. To construct these models the BioConductor library function "glm" was used with the "family" parameter set to "binomial" [Hastie and Pregibon, 1992][Venables and Ripley, 2002].

5.5.7 Estimates of performance characteristics

For clinical applications, classification performance is generally reported and compared in terms of sensitivity and specificity or related classical diagnostic terms [Pepe et al., 2001]. Diagnostic performance, e.g. sensitivity and specificity is estimated for many of the candidate biomarkers and these values are

used throughout this report. These values are derived from the number of actual specimens in a disease (or normal) class and the number of tests (assays, statistical or otherwise) that are positive (or negative) as defined in Table 5.4: Using these defined terms, performance characteristics were calculated as shown

Table 5.4: Clinical descriptors of test performance assuming a two-class phenotype case (e.g. neoplasia vs. non.neoplasia) and where classifier refers generally to any discrimination technique or technology e.g. clinical assay, multivariate model, etc.

Result	Definition
True Positive	classifier and phenotype AGREE where phenotype is positive
False Positive	classifier and phenotype DISAGREE where phenotype is negative
True Negative	classifier and phenotype AGREE where phenotype is negative
False Negative	classifier and phenotype DISAGREE where phenotype is positive

in Table 5.5. With the important exception of "hypothesis testing", many of

Table 5.5: Formulas for commonly used assay performance terms are shown. The following abbreviations are used for all calculations: true positives (TP), false positives (FP), true negatives (TN), false negatives (FN), positive predictive value (PPV), negative predictive value (NPV), likelihood ratio positive (LRP), and likelihood ratio negative (LRN).

Term	Formula
sensitivity	TP / (TP + FN)
specificity	TN / (FP + TN)
PPV	TP / (TP + FP)
NPV	TN / (TN + FN)
LRP	sensitivity / (1 - specificity)
LRN	(1 - sensitivity) / specificity

the performance metrics reported here were based on calculations made in precisely (or effectively) the same data set as was used to discover the predictor in the first place. Such performance metric estimates will therefore potentially overestimate, perhaps to an unreasonably large degree, the performance characteristics of such tests in future tissue samples sourced from independent clinical

populations. To improve the generalisability of the estimates a modified jack-knife re-sampling technique was used to calculate a less biased value for each characteristic [Hastie et al., 2001].

Importantly, the key conclusions of this research are derived from hypotheses tested in independent clinical samples and do not suffer from this reporting weakness. Nevertheless, given the usual effects of sample size, etc. performance characteristic estimates are still reported using estimated confidence intervals where useful.

5.5.8 Receiver operator characteristic curves and D-Value

As discussed in 3.2.3, the relative costs for diagnosis are often not equal for each predicted phenotype. For example the costs associated with missing a disease may greatly outweigh the costs of a false-positive diagnosis. Alternatively, the follow-up costs or downstream procedural risks associated with a positive diagnosis may influence a rational health care decision to minimise false negative interpretations. Thus, there may be an advantage to understanding the dynamic relationship between diagnostic performance measures (e.g. sensitivity) and adjusting the threshold for a positive diagnosis. The use of receive operator characteristic (ROC) curves can improve this understanding [Pepe et al., 2001].

For select candidates, threshold-response relationships for classification are illustrated as using ROC plots or curves. For routine analyses, however, ROC results are often summarized through the convenient D statistic, or 'effectiveness parameter' described in Saunders [2006]. Assuming a normally distributed biomarker, the D statistic is related to the area under the curve (AUC) of an ROC plot by $\Phi(D/\sqrt{2})$, where Φ is the Gaussian distribution function.

One advantage of the D statistic is the ability to conveniently estimate confidence intervals which are not conveniently estimated for ROC curves. The D value is interpretable as a measure of the disease impact on a biomarker as a proportion of the variation across the test population [Saunders, 2006]. Further,

under assumptions of normality, D can be related precisely to the ROC curve exactly *at that point were sensitivity = specificity*:

$$\text{sensitivity} = \text{specificity} = \Phi[\frac{D}{2}]$$

Thus the D value is often used herein where a full ROC plot is unnecessary. Finally, a Bayesian estimate of the D parameter has been implemented which is often used to estimate a 95% confidence interval for D and sensitivity and specificity[Saunders, 2008]. Sensitivity and specificity estimates in this context assume that biomarker values are normally distributed, but even if this assumption is not valid, this metric provides an objective metric for comparing multiple biomarkers or biomarker panels.

5.5.9 Tissue specific expression patterns

Discovery methods using gene expression data often yield numerous candidates, many of which are not suitable for commercial products because they involve subtle gene expression differences that would be difficult to detect in laboratory practice. Pepe et al. note that the 'ideal' biomarker is detectable in tumor tissue but not detectable (at all) in non-tumour tissue [Pepe et al., 2001]. Historically (and until this point), this ideal has been viewed as unlikely to be achieved. Screening and diagnostic tests, which are two particular uses of these biomarkers, are well-characterised for their capacity to predict likelihood of the target lesion being present. Nonetheless, to bias discovery toward candidates that most closely behave as ideal biomarkers, the author has developed an analysis method which aims to enrich the candidates for biomarkers whose qualitative absence or presence measurement is highly sensitive and specific for the phenotype of interest. Such candidates would most closely meet the Pepe criterion and such biomarkers would be preferred discriminators, i.e. predictors of likelihood. This method attempts to select candidates that show a prototypical "turned-on" or "turned-off" pattern relative to an estimate of the background/noise expression across the microarray. Such RNA transcripts may a) correlate with downstream

translated proteins that have diagnostic potential; or b) predict upstream genomic changes (e.g. methylation status) that could be used diagnostically. This focus on qualitative rather than quantitative outcomes could simplify the product development process for such biomarkers.

The method is based on the assumption that the pool of extracted RNA species in any given tissue (e.g. colorectal mucosa) will specifically bind to a relatively small subset of the full set of probesets on a microarray designed to measure the whole genome. A consequence of this assumption is that *most* probesets on a full human genome microarray experiment will *not* exhibit specific, high-intensity signals.

To approximate the background or "non-specific binding" across the entire experiment, a gene expression level approximately equal to the value of lowest 30% quantile of the ranked mean values was determined. This quantile threshold can be arbitrarily set to some level below which one could reasonably assume that the signals do not represent specific (i.e. higher than background) RNA transcript binding. Thus this gene expression level equal to the 30% highest ranked expression value is used as the threshold for qualitatively determining a probeset to be "on" or "off" by being above or below this cutoff, respectively.

For this work a range of quantile cutoffs (5%-40%) was explored and a 30% threshold was found to yield manageable probeset list sizes for subsequent validation.

Conversely, there is a tacit assumption that probesets are a) expressed above this theoretical threshold level and b) expressed at (statistically significant) elevated levels in the tumour specimens may be a *tumour specific* candidate biomarker. Also, a third criterion based on "fold-change" thresholds can also be conveniently applied to further emphasize the concept of absolute expression increases in a putatively "ON" probeset.

Given the assumption of low background binding for a sizable fraction of the measured probesets, this method was only used in the large GeneLogic discovery data. To construct a filter for hypothetically "turned on" biomarker in these

data, the mean expression level for all 44,928 probesets was estimated across the full set of 454 tissues. These 44,928 mean values were then ranked and the expression value equivalent to the 30th percentile across the data set was determined. This arbitrary threshold was chosen as a conservative estimate below which that proportion of RNA species in a given specimen should exhibit low concentration effectively equivalent to transcriptional silence. Thus, this threshold represents a conservative upper bound estimate of non-specific, or background, expression. Figure 5.1 shows the distribution of chip intensity values for all probesets in all tissues. This plot suggests that the majority of probesets are normally distributed around a low level of gene expression across all microarrays. This distribution of intensity values is consistent with a population of probesets exhibiting background, or non-specific "noise" binding.

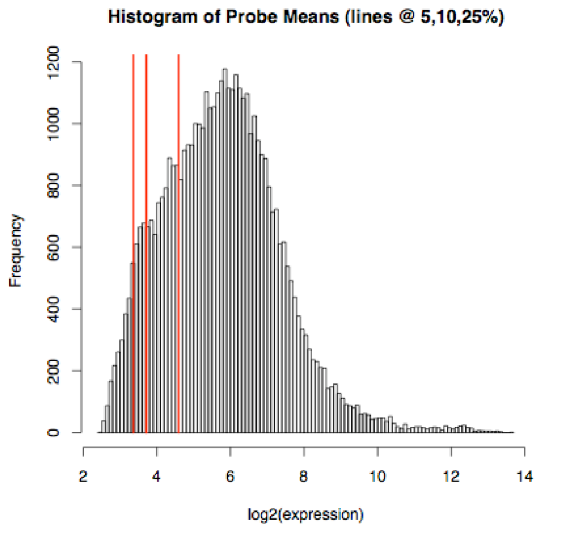


Figure 5.1: Histogram of 44,928 mean probeset intensity values (log base 2) averaged over all 454 chips. The 5%, 10% and 25% threshold values are indicated by red. A majority of probesets are approximately normally distributed around approximately 2^6.

5.5.10 Gene set enrichment analysis

Gene set enrichment analysis (GSEA) involves testing the hypothesis that a defined set of genes is differentially expressed *in concert* between two or more phenotypes of interest [Subramanian et al., 2005]. For example, GSEA was used in the work to test whether particular sets of genes for a given transcriptional pathway, such as the Wnt pathway, are differentially expressed *as a group* between neoplastic and non-neoplastic observations. To explore this question, the GSA library for R described in Efron and Tibshirani [2006] was used.

GSEA requires *a priori* defined gene sets to test for group-wise differential expression. Publicly available gene lists are available for this purpose and for this work the Kyoto (KEGG) database was used: BioConductor.org, package version 2.2: created Friday, Apr. 2 09:54:29 2008 [Liu et al., 2008a][Kanehisa et al., 2008].

For comparison, a manually curated list of Wnt targets was assembled based on R. Nusses' Wnt Homepage (See `http://www.stanford.edu/%7ernusse/wntwindow.html`) [Nusse, 2008] and also the literature review described in Table 2.1. Interestingly, this manually curated list did *not* strongly overlap with the publicly available Wnt list published by KEGG (see 7.4). Consequently, this manually curated list was included as an "experimental" Wnt list in these studies. A complete list of the gene sets used for each of the GSEA experiments and a list of genes used in the manually curated Wnt list are provided in the Appendix in Table D.2 and Table D.3, respectively.

5.5.11 K-nearest neighbor clustering

K-Nearest Neighbor (KNN) is a clustering metric whereby observations (tissues) are projected into a multidimensional space defined by some feature space and then each observations is compared with neighboring observations. A KNN implementation was designed to test all-subsets of candidate genes from the differential discovery data set. The goal of this analysis was to find the best

p-dimensional gene set capable of clustering tissue data by phenotype (i.e. neoplasia vs. non-neoplastic control data).

The algorithm is shown in Algorithm 1.

For the analysis of the differential discovery data, a range of k parameter values ($1 \leq k \leq 5$) were tested and $k = 3$ was found to yield reasonable, if perhaps conservative, results.

When applied to the relatively small differential display RT-PCR data of 67 primer targets measured in 71 observations (tissues), all combinations of up to four genes at a time were explored by all-subsets testing in reasonable computational time (hours). Code for the all-subsets algorithm was written in C++ and R.

5.5.12 Genetic algorithm for KNN

To explore higher dimensional sets using five or more RT-PCR gene targets, the KNN algorithm was wrapped into a genetic algorithm designed and coded in C++ by the author. The algorithm first created a seed search space of 10 pools of 5,000 p-length vectors comprised of randomly assigned RT-PCR genes. Next, the top 100 vectors containing the highest KNN scores (see KNN details above) were injected into each pool for the next round of searching (the remaining 4,900 vectors otherwise carried over from each pool). Further, for each of the 50,000 vectors there was a small (0.5%) chance of a random target change at any position for each round of search.

Using this algorithm, vectors comprising up to 5, 8, 10, 12, and 15 biomarker candidates were evaluated through twenty generations.

5.5.13 Principal components analysis

Principal components analysis was used extensively in this research to explore relationships between phenotype and global variance.

Algorithm 1 Subset selection algorithm using a brute-force KNN analysis.

Repeat

1. for each combination of P genes taken p genes at a time:

   (a) Repeat

      i. For each observation x_i where $i \in 1, \ldots, N$:

         A. Project each x_i into p-space using the p genes

         B. Create a distance matrix between all observations, using any metric of choice. For this implementation we employed the usual Euclidean distance metric:

$$\begin{gathered}\text{Given point } A = (a_1, a_2, \ldots, a_p) \\ \text{and point } Z = (z_1, z_2, \ldots, z_p) \text{ in } p \text{ space.} \\ \Delta(A, Z) = \sum_i^p \sqrt{(a_i - z_i)^2}.\end{gathered}$$

         C. Choose the k observations nearest to x_i based on this distance matrix.

         D. Assign an experimental classification $\hat{C}$ to each x_i based on the true classification of the k observations nearest to x_i. We use a unanimous decision rule for this experimental assignment. Lack of concurrence results in an 'unknown class' designation.

      ii. Choose next i.

   (b) Calculate a score for this p combination of genes by:

$$\text{SCORE} = \sum_i^N (\text{Correct classifications})$$

2. Choose next p combination.

The combination of p genes with the best score is selected.

To calculate the principal components of an N by p matrix $\mathbf{X}$, the following routine was used (in R):

1. Mean center $\mathbf{X}$ by shifting each column $\mathbf{x}_i$ such that $\mu_i = 0$.

2. Calculate the singular value decomposition of (or eigendecomposition of $\mathbf{X}^t\mathbf{X}$) as follows

   $$\text{where } \mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^t,$$

   $\mathbf{v_i}, i \in \{1, \cdots, p\}$ is the ith principal component.

The principal components, also called the Karhunen-Loeve directions, are the p vectors in the direction of decreasing variance of the N vectors in p space from the data $\mathbf{X}$.

The first principal component, $\mathbf{v_1}$, is the p-length vector in the direction of highest variance across the squared data and each subsequent vector is orthogonal to all others in decreasing order of variance in that direction. Thus, the data projection $\mathbf{z} = \mathbf{X}\mathbf{v_1}$, has the highest variance of all linear combinations of the columns of $\mathbf{X}$ and

$$\text{var}(\mathbf{z_1}) = \text{var}(\mathbf{X}\mathbf{v_1}) = \frac{\lambda_1^2}{N}$$

decreases as i increases such that $\mathbf{z}_{i=p}$ has the lowest variance and λ_i is the ith eigenvalue.

5.5.14 Supervised principal components analysis

Supervised principal components analysis was used to visualize and explore the high dimensional structure of expression differences between phenotypes. Supervised PCA (sPCA) is similar to traditional principal components analysis but uses only a subset of the features/genes (usually selected by some univariate or multivariate means) to derive the principal components [Bair et al., 2006].

To perform sPCA, first a subset of the data $\mathbf{X}$ with a reduced number of features,p^* is extracted, usually selected by univariate means e.g. by t testing to

find differentially expressed probesets. The usual PCA analysis is then applied as described above to the subset $\mathbf{X}^*$, which is N by p^*.

sPCA was usefully applied in this work, for example, to understand the nature of gene expression patterns along the large intestine.

5.6 Conclusions

This chapter described the two data sets used in this research for generating biomarker hypotheses of differential gene expression between neoplastic and non-neoplastic phenotypes. Collectively these data are referred to throughout this work as the "discovery" data. A third set of data is also described that was then used to test these biomarker hypotheses. This set of "validation" data was generated using a custom microarray designed by the author and manufactured by Affymetrix.

All clinical specimens and relevant laboratory methods were presented in this chapter.

Finally, the key statistical and analytical methods used throughout this research were described. In particular, a novel method of gene expression analysis was introduced which is motivated by the desire to filter differentially expressed probesets to yield probesets that may exhibit a "turned-on" or "turned-off" expression profile in one phenotype but not the other. Such probesets are sometimes described herein as neoplasia-specific.

Chapter 6

Normal Gene Expression

6.1 Aim

This thesis tests the hypothesis that gene expression differs between neoplastic tissues and non-neoplastic controls. An understanding of gene expression in the normal state is inherent to testing this hypothesis.

This chapter aims to explore patterns of gene expression along the longitudinal length of the large intestine. To mitigate the impact of disease-related gene expression changes, a large set of transcripts derived from histologically-normal tissue specimens were analyzed. Both univariate and multivariate methodologies were applied to explore these data.

6.2 Introduction

To date little is known about how much variation occurs in normal tissues and whether the magnitude of such variation poses problems for comparing neoplastic and non-neoplastic tissues. Furthermore, the colorectum is a long organ with changing physiology, differing ontology and differing exposure to extrinsic factors such as dietary contents and microflora along its length. These potentially confounding factors introduce the possibility of large gene expression variability

in normal tissues and require examination prior to experimenting with diseased tissues.

The advent of gene expression profiling has led to an improved understanding of intestinal mucosa development. For example, the regulation of transcription factors involved in producing and maintaining the radial-axis balance from the crypt base to the lumen and those giving rise to epithelial cell differentiation are now better understood as a result of microarray gene expression analysis [Peifer, 2002, Traber, 1999]. Similarly, understanding of the developmentally programmed genetic events within the embryonic gut has improved, especially those molecular control mechanisms responsible for regional epithelium differences between the small intestine and colon [de Santa Barbara et al., 2003, Park et al., 2005]. On the other hand, little is known about the proximal-distal gene expression variation along the longitudinal axis of the colorectum in either the neoplastic or non-neoplastic setting [Bates et al., 2002]. Epidemiological studies of colorectal adenocarcinoma suggest support for variable incidence, histopathology, and prognosis between proximal and distal tumors [Bonithon-Kopp and Benhamiche, 1999, Bufill, 1990, Deng et al., 2002, Distler and Holt, 1997]. Thus an understanding of location-specific variation could provide valuable insight into those diseases that have characteristic distribution patterns along the colorectum, including colorectal cancer [Birkenkamp-Demtroder et al., 2005, Caldero et al., 1989, Garcia-Hirschfeld Garcia et al., 1999].

The large intestine is divided into six anatomical regions starting just beyond the terminal region of the ileum: the cecum; the ascending colon; the transverse colon; the descending colon; the sigmoid colon; and the rectum. Alternatively, these segments may be grouped to divide the large intestine into a two region model comprising the proximal and distal large intestine. The proximal ("right") region is generally taken to include the cecum, ascending colon, and the transverse colon while the distal ("left") region includes the splenic flexure, the descending colon, the sigmoid colon and the rectum. This division is supported by the distinct embryonic ontogenesis of these regions whose junction is two thirds along the transverse colon and also by the distinct arterial

supply to each region. While the proximal large intestine develops from the embryonic midgut and is supplied by the superior mesenteric artery, the distal large intestine forms from the embryonic hindgut and is supplied by the inferior mesenteric artery [Babyatsky and Podolsky, 2003]. A comprehensive of review of proximal/distal differences are provided in Iacopetta [2002].

The longitudinal nature of the large intestine along the proximal-distal axis provides a relatively unique opportunity for constructing a whole organ map of gene expression. Previous research suggests that there is a clear distinction between the gene expression patterns of proximal colonic tissues and distal colorectal tissues [Glebov et al., 2003, Birkenkamp-Demtroder et al., 2005, Komuro et al., 2005]. While these findings support a broad model of gene expression difference, there have been no studies to explore the detailed nature of expression gradients of such genes. Given the interesting embryology related to the midgut and hindgut junction near the splenic flexure during embryogenesis, the question is raised: do differentially-expressed genes exhibit an abrupt expression schism between the midgut and hindgut derived tissues or does expression follow a gentle gradient along the proximal-distal axis?

To explore this question, a formal hypothesis which tests whether a more complex multi-segment model statistically improves the description of the multi-segment gene expression relative to a simple proximal vs. distal model was tested. Such hypotheses were constructed and tested for each probeset that exhibits statistically significant differential expression between the caecum and the rectum.

Exploration of these patterns in non-neoplastic tissues may improve the understanding of gene expression variation in healthy normal adults without the added complexity of neoplasia-related gene expression changes. Expression profile "maps" were built that identify individual genes whose expression appears to be location dependent and the nature of multi-gene expression variance longitudinally along the colon is also described. Linear models were applied to these maps to compare the embryology-consistent proximal vs. distal two-region model with a more gradual model based on continuously variable expression be-

tween the cecum proximally and rectum distally. Such gene expression maps of the normal adult colon will provide a foundation for improved understanding of gene expression variation in both the normal and diseased state.

6.3 Results

6.3.1 Gene expression data

To explore gene expression along the non-neoplastic colon, Affymetrix (Santa Clara, CA USA) GeneChip(R) oligonucleotide microarrays such as those described in Lipshutz et al. [1999] were analyzed. The data are two independent Affymetrix (Santa Clara, CA USA) Human Genome 133 GeneChip data sets: a large commercial microarray database of HGU-133 A&B chip data for "discovery", and a smaller HGU-133 Plus 2.0 microarray data set generated by the author for "testing". The larger data set was purchased to identify gene expression patterns and the independently derived second expression set was used to test these patterns.

These data are further described in Chapter 5.

Discovery data

To construct the discovery set, 184 GeneChips hybridised to cRNA from non-diseased tissues meeting inclusion and quality assurance criteria were used for hypothesis generation. The tissues comprised segment subsets as follows: 29 cecum, 45 ascending, 13 descending, 54 sigmoid, and 43 rectum. For each tissue, 44,928 probe sets were background corrected and normalised using RMA pre-processing. The theoretical juncture between the proximal and distal colon is approximately two-thirds the length of the transverse colon measured from the hepatic flexure [Babyatsky and Podolsky, 2003]. As sample data were not specific for distance along the transverse colon, these tissues were excluded from the discovery analysis.

Test data

To construct the validation, or "test", data set, 19 HG U133 Plus2.0 GeneChips were hybridised to labeled cRNA prepared from 8 proximal tissue specimens and 11 distal specimens from the Repatriation General Hospital (Adelaide, SA) tissue bank. Due to stringent quality control parameters for tissue and GeneChip acceptability, this validation data set did not include sufficient tissues to explore multiple segment models. Each microarray measured transcript expression for 54,675 probe sets.

6.3.2 Gene variation along the colon: univariate analyses

To explore the "natural" dividing point between the anatomical segments of the colon, the absolute number of significant probeset expression differences was measured by modified t test when the hypothetical "divide" was moved stepwise from caecum to rectum [Smyth, 2005]. Figure 6.1 shows the number of probesets that were differentially expressed for each inter-segment divide. The maximum number of probeset differences, 206, occurs when the proximal and distal regions are divided between the ascending and descending segments which is slightly higher than the number of differences between the descending and sigmoid segments. Interestingly, there were many fewer differential genes between the other colon segments and the rectum. As the dividing point between the ascending and descending colon is consistent with both the accepted understanding of embryonic development and the usual separation of the proximal and distal segments, the following comparison of proximal and distal tissues were based on this division.

A total of 206 probesets, corresponding to approximately 154 presumed gene symbols, were differentially expressed higher in the proximal or distal colorectal samples compared to the complementary region (Bonferroni corrected $P < 0.05$). Of these 206 probesets, 31 (16.5 %) were also differentially expressed in the validation data with a significant difference (31/206, $P << 10^(- 5)$ by Monte

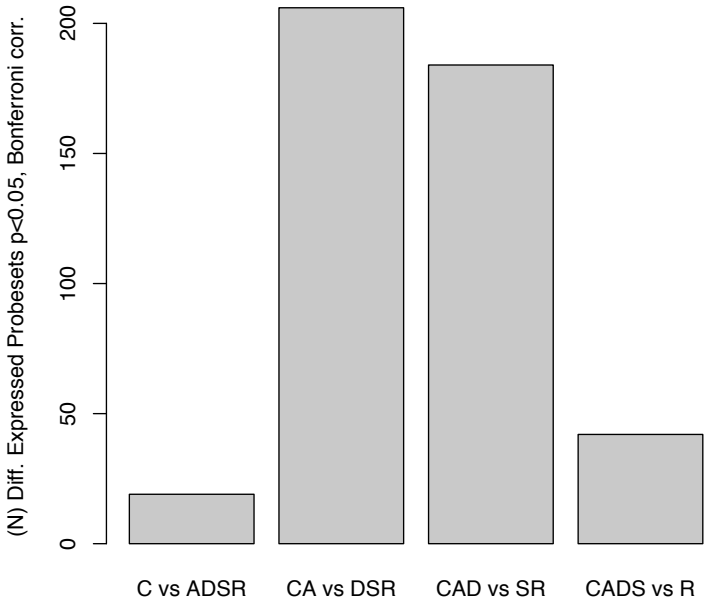


Figure 6.1: The number of differentially expressed measured by t test ($P <= 0.05$) is shown between each possible proximal-distal dividing point along the colorectum. Segments are designated caecum (C), ascending colon (A), descending colon (D), Sigmoid colon (S), and rectum (R). The maximum number of differential genes is observed using a break-point between ascending and descending colon, however there nearly as many differential genes if one uses a break-point between the descending and sigmoid colon.

Carlo estimation).

To further explore differential expression in the discovery set, we identified those transcripts that were different between the most terminal ends of the large bowel. A total of 115 probesets were differentially expressed between tissues selected only from the caecum ($N = 29$) and the rectum ($N = 43$). 102 (89%) of these probesets were included in the 206 probesets differing between proximal and distal colon described above. In this subset, 28 probesets (24.3%) were likewise differentially expressed in the rectum vs. the cecum in the validation data (28/115, $P << 10^{(} - 5$) by Monte Carlo estimation). All 28 of these consistent probesets were included in the 31 consistent probesets between the distal and proximal regions.

Differentially expressed probesets and difference statistics for probesets with elevated expression in proximal and distal tissues are shown in Appended Tables 4.1, p.262 and 4.2, p. 263 respectively.

An analysis for differential expression was also made for all five inter-segment transitions in order from the cecum to the rectum (i.e. cecum vs. ascending, ascending vs. descending, etc.). No transcript was statistically differentially expressed between any two adjoining segments (limma t-test; $P > 0.05$).

To explore the nature of these gene transcript expression changes, we built and compared linear models fitted to the expression data based on location for each tissue sample. Two linear models of univariate probeset expression were compared for each of the 115 probesets differentially expressed between the two terminal segments of the large intestine, the cecum and rectum. In particular, we queried whether the expression of those transcripts that were differentially expressed between these terminal segments were better explained (in terms of residual fit) by a simple two-segment model or by the more descriptive five-segment model.

Of the 115 differentially expressed probesets, the analysis failed to reject the null hypothesis that a complex model does not significantly improve model fit to the observed gene expression data for 65 (57%) of cases (F-test, $p > 0.05$). Thus, more than half of these differentially expressed transcripts along the colon are satisfactorily modeled by the two segment expression model whereby expression is dichotomous and defined by either proximal vs. distal location. The most differentially expressed probeset between the cecum and rectum is the designed to hybridise against the *PRAC* gene transcript. A comparison of the two-segment and multi-segment models for this transcript are shown in Figure 6.2, which is typical of other genes in this proximal vs. distal category. This expression pattern for PRAC was also confirmed by RT-PCR analysis as shown in Figure 6.3.

For the remaining 50 (43%) probesets, the null hypothesis was rejected ($p < 0.05$) which suggested that a five factor model dependent on segment location in

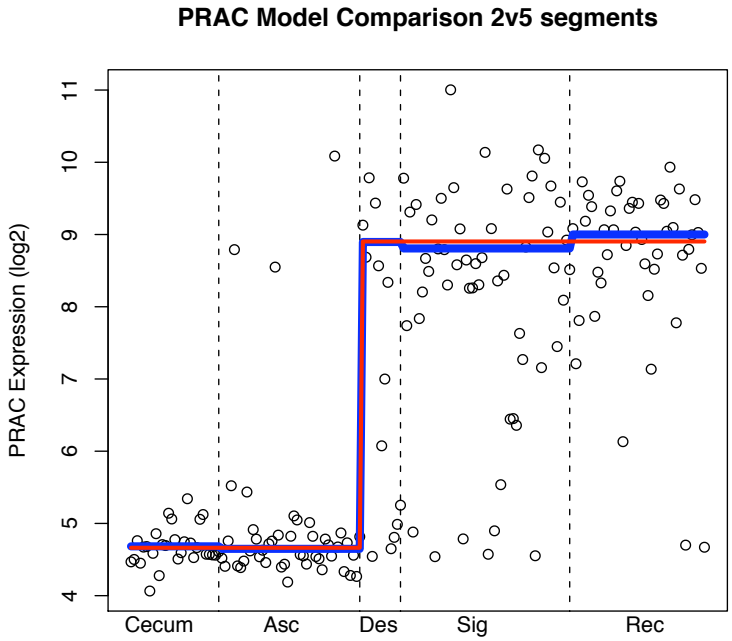


Figure 6.2: Gene expression measurements for the *PRAC* gene grouped by anatomical segment illustrating the dichotomous/binary pattern that is the dominant pattern of transcript expression along the proximal-distal axis. Shown in red is a two segment model fit to these data while a five segment model is shown in blue. There is no significant improvement of fit using the more complex five segment model. Note that the ordering *within* each segment is essentially random and no further data are available regarding intra-segment distances.

fact improves the predictive effectiveness of such transcripts' expression along the proximal-distal axis in a significant manner. Inspection of these models confirms that most probeset levels are monotonic-increasing or monotonic-decreasing in tissues progressing along the large intestine. 41 (82%) of the 50 multi-segment models exhibited a gradual transcript level increase across the colon while only 9 models (18%) indicate a gradual decrease from proximal to distal expression. The model for homeobox gene B13 (*HOXB13*) is significantly improved with the five segment model compared to the two segment model as illustrated in Figure 6.4.

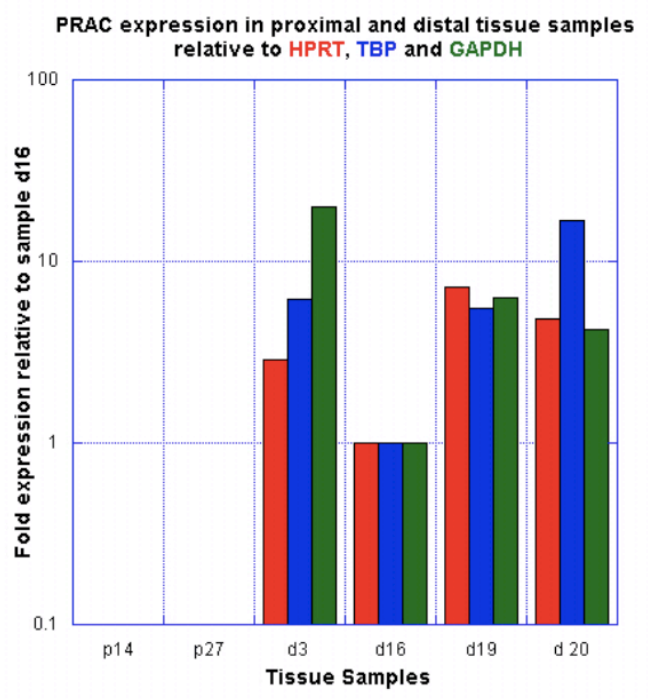


Figure 6.3: Gene expression measurements of *PRAC* using real-time PCR. *PRAC* concentrations were measured in six tissues (2 proximal and 4 distal). Each measurement was normalised against three "housekeeping" genes which are widely used in the literature: *HPRT* (RED), *TBP* (BLUE) and *GAPDH* (GREEN). For each tissue sample, normalised concentrations of *PRAC* are comparable across normalization methods. There is no measurable *PRAC* mRNA in the two proximal tissues while distal tissues exhibit a range of transcript expression. These data confirm the microarray data results and show increased expression in the distal colorectum.

6.3.3 Patterns of gene expression along the colon

In addition to analyses of individual gene changes along the colon, we used multivariate analytical techniques to explore patterns of gene changes along the proximal-distal axis.

PCA and supervised PCA

The full 44,928 probesets of the "discovery" data set were analyzed using PCA. The first two dimensions of this analysis are shown in Figure 6.5, p. 112. Inspection of this two-dimensional perspective yields no obvious structure within the

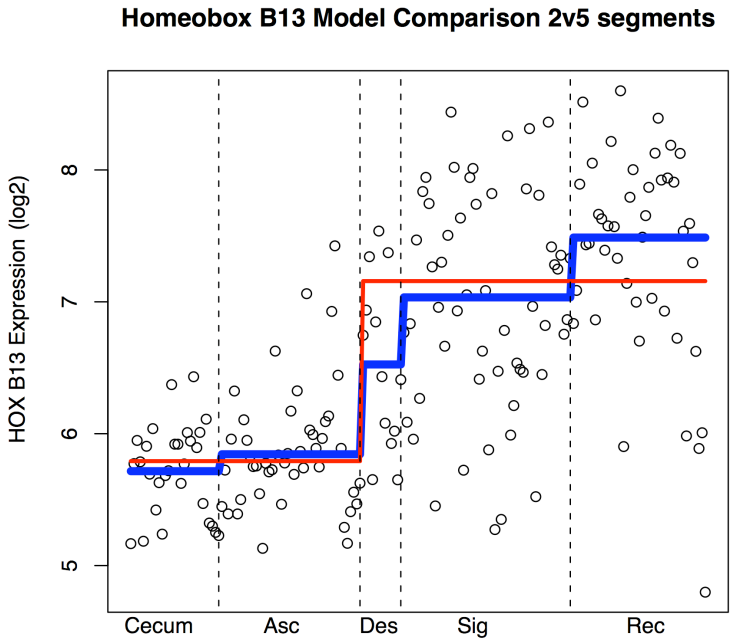


Figure 6.4: Gene expression measurements for the *HOXB13* gene grouped by anatomical segment illustrating the second pattern observed along the proximal-distal axis: a gradual change from segment to segment. Shown in red is a two segment model fit to these data while a five segment model is shown in blue. Unlike the *PRAC* data shown above, these gene expression data show a significantly improved fit to them five segment model and the null hypothesis is rejected. Note that the ordering *within* each segment is essentially random and no further data are available regarding intra-segment distances.

data. This analysis suggests that the major sources of gene expression variation (i.e. the first two principal components) measured across all genes between the tissue samples does not correlate with tissue location.

Nevertheless, while tissue location may not correlate with the major directions of variance in these data, at least a subset of probesets are differentially expressed between the proximal and distal colorectum. To therefore explore the nature of regional expression further, a supervised PCA (sPCA) was also applied to the data. sPCA is similar to traditional principal components analysis but uses only a subset of the features/genes (usually selected by some univariate means) to derive the principal components (see section 5.5.14 for further details). The subset of probesets differentially expressed between the cecum and rectum as

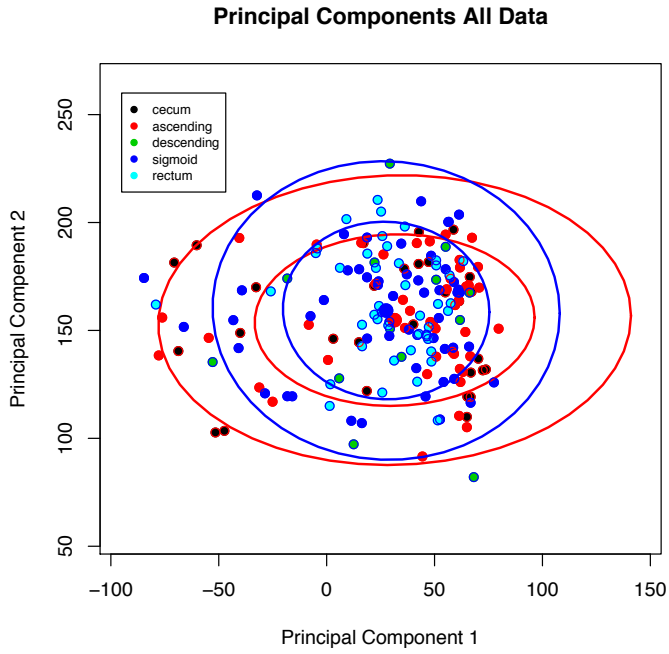


Figure 6.5: Principal component analysis using all probesets across all tissues which are color coded by anatomical segment: caecum (black), ascending (red), descending (green), sigmoid (dark blue) and rectum (light blue.) This plot shows that there is no obvious multi-phenotype clustering in these data associated with anatomical location from a genome-wide perspective.

described above were used for sPCA: a reduced data matrix of all 184 normal tissues was constructed with just the top 115 probesets differentially expressed between the cecum and rectum. PCA was then performed using this feature specific data and the 184 tissues were again visualized along just the first two principal components, shown in Figure 6.6. Inspection of Figure 6.6 indicates that there are two broad populations within these tissues corresponding approximately to the proximal vs. distal divide. By reducing the dimensionality of this projection to just a single first component as shown in Figure 6.7 and Figure 6.8, the proximal vs. distal relationship became clear. There is strong overlap between the sigmoid colon and rectum segments at the distal end and between the segments of cecum and ascending colon at the proximal end.

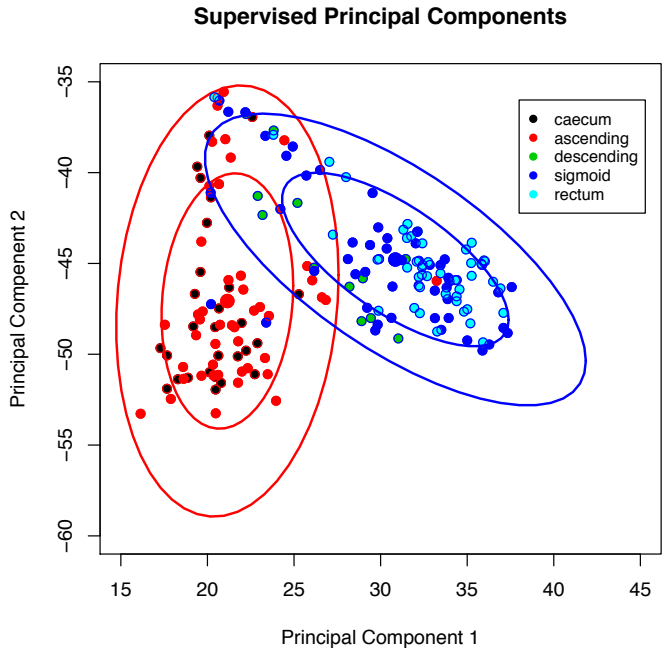


Figure 6.6: Supervised principal component analysis using *only* the 115 probesets which are differentially expressed between the caecal and rectal tissues. Individual tissue observations are color coded by anatomical segment: caecum (black), ascending (red), descending (green), sigmoid (dark blue) and rectum (light blue.) This plot demonstrates that colorectal tissue location correlates strongly with the two observed clusters.

6.4 Discussion

6.4.1 A map of differential gene expression along the colon

These data show that tissue location is not the dominant source of variation among these 184 non-diseased colorectal tissues. Of 44,928 probesets measured, only 206 exhibit significant difference of gene expression means between proximal and distal tissues. These 206 probesets correspond to approximately 154 unique gene targets that are differentially expressed between the normal proximal and normal distal large intestine regions in human adults. A subset of 115 probesets (89% common to the proximal vs. distal list) is likewise differentially expressed between the terminal colorectal segments of the cecum and rectum. Interestingly, no transcripts were observed to be differently expressed between

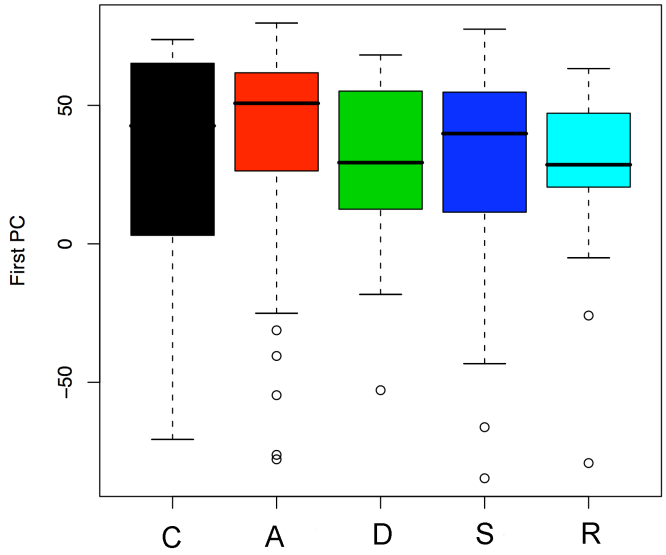


Figure 6.7: Boxplot of first principal component values taken from the genome-wide gene expression data grouped by segment: caecum (C,black), ascending (A,red), descending (D,green), sigmoid (S,dark blue), and rectum (R,light blue). An analysis of this first principal component using all 44,928 probesets shows that the first primary component of variance is not correlated with anatomical location.

any two adjacent segments.

To estimate the validity of these findings, the expression change of these differential probesets were validated in an independent set of microarray data. Thirty-one (31) of the 206 differentially expressed probesets in the initial discovery data set of 184 colorectal tissue samples were also differentially expressed in the test data of 19 specimens.

Nearly all (28/31, 90%) of these "confirmed" transcripts were likewise differentially expressed between the two terminal segments of the cecum and rectum.

Some of the probesets described herein are designed to hybridise to gene transcripts that were previously identified to be differentially expressed by microarray analysis using a variety of cDNA and oligonucleotide microarrays [Glebov

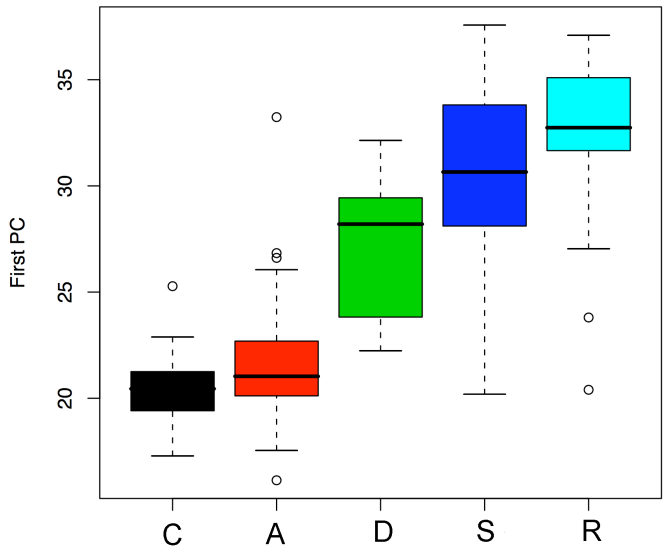


Figure 6.8: Boxplot of first principal component values taken from the limited set of gene expression data using only differentially expressed probesets between the caecum and rectum. The data are grouped by segment: caecum (C,black), ascending (A,red), descending (D,green), sigmoid (S,dark blue), and rectum (R,light blue). These data suggest that the proximal-distal differential expression pattern is stronger than the inter-segment differential expression and that there is a general trend which correlates with anatomical segment moving distally along the colorectum.

et al., 2003, Birkenkamp-Demtroder et al., 2005, Komuro et al., 2005]. Five of the gene targets of differential probesets described here were previously identified in two or more of these earlier studies, including: *HOXB13*, *NR1H4*, *S100P*, *SCNN1B*, and *SIAT4C*. Each of these probesets were also shown to be statistically different (i.e. *HOXB13*, *SIAT4C*: $P < 0.065$), in the validation data set. An additional 33 probeset target genes of the 206 probesets were previously identified to be differentially expressed along the colon in at least one of these earlier studies.

An additional 28 probesets that were differential in both the discovery data and the independent test data but were not reported in the previous reports were identified. In total, 57 of 154 (37%) gene targets corresponding to the 206

probesets were confirmed to be differentially expressed between the proximal and distal from the validation set. The agreement of this work with earlier studies and with the independent validation set adds credibility to the results, especially given the potential for concern about microarray reproducibility between and within data collection platforms [Miklos and Maleszka, 2004].

6.4.2 Expression patterns of selected genes

The most significantly differential probeset observed in these discovery data was against the gene transcript for *PRAC*, previously described as specifically expressed in prostate, the distal colon and rectum [Liu et al., 2001]. These data agree with the earlier findings that the probeset for *PRAC* is highly expressed in the distal colon relative to the proximal tissues. This observation was confirmed by RT-PCR. Further, *PRAC* appears to be expressed in a low-high pattern along the colon with a sharp expression change occurring between the ascending and descending colorectal specimens.

Eight probesets corresponding to seven *HOX* genes were found to be differentially expressed between the proximal and distal colon. The 39 members of the mammalian homeobox gene family consist of highly conserved transcription factors that specify the identity of body segments along the anterior-posterior axis of the developing embryo [Hostikka and Capecchi, 1998, Kosaki et al., 2002]. The four groups of HOX gene prologues are expressed in an anterior to posterior sequence, for e.g. from *HOXA1* to *HOXB13* [Montgomery et al., 1999]. The expression patterns for these eight probesets is consistent with the expected pattern: lower numbered HOX genes are expressed higher in the proximal tissues (*HOXD3*, *HOXD4*, *HOXB6*, *HOXC6* and *HOXA9*), while the higher named genes are more expressed in the distal colon (*HOXB13* and *HOXD13*). Elevated expression of *HOXB13* in the distal colon was confirmed by RT-PCR (see Figures 4.3, p. 264). These results are also consistent with examples of specific HOX expression in the literature, such as studies that demonstrate *HOXD13* involvement in the development of the anal sphincter in mice [Kondo et al.,

1996].

There was, however, conspicuous absence in these findings of some gene transcripts that have been previously shown to be differentially expressed along the proximal-distal axis. These data do not demonstrate a significant expression gradient for the caudal homeobox genes *CDX1* or *CDX2*, transcription factors that have been shown to be involved in intestine pattern development across a range of vertebrates [Chalmers et al., 2000, James et al., 1994, Silberg et al., 2000]. In particular, *CDX2* is considered to play a role in maintaining the colonic phenotype in the adult colon and was shown to be present at relatively high concentrations in the proximal colon but absent in the distal colon [James et al., 1994, Silberg et al., 2000]. Neither statistical analysis nor visual inspection of probeset expression for this gene suggest differential expression along the colon in these data (data not shown). Analysis by RT-PCR of a subset of RNA samples from the validation set supported the array data in that expression of *CDX2* in the distal colon was equivalent to or greater than in proximal samples (see Figures 4.3, p. 264).

Significant differential transcript expression was observed for a number of the solute-carrier transport genes that can be rationalized based on accepted understanding of colorectal physiology. While probeset expression for *SLC2A10*, *SLC13A2*, and *SLC28A2* are higher in the distal colon, the solute carrier family members *SLC9A3*, *SLC14A2*, *SLC16A1*, *SLC20A1*, *SLC23A3*, and *SLC37A2* are higher in the proximal tissues. These data support the findings of Glebov et al., including for the Na-dependent dicarboxylic acid transporter member 2 (*SLC13A2*) which is elevated distally and for the monocarboxylic acid transporter family member 1 (*SLC16A1*, alias *MCT1*) which is higher in the proximal tissues [Glebov et al., 2003]. This expression of *SLC16A1/MCT1* is consistent with evidence that the short chain fatty acid butyrate, which is most abundant in the proximal gut [Macfarlane et al., 1992], may regulate *SLC16A1/MCT1* expression by both transcriptional control and by transcript stabilization [Cuff et al., 2002].

These results show that probesets against all three of the five members of the

chromosome 7q22 cluster of membrane-bound mucins previously believed to be expressed in colon, *MUC11*, *MUC12* and *MUC17*, are differentially expressed at elevated levels in the distal gut [Byrd and Bresalier, 2004, Williams et al., 1999, Gum et al., 2002]. We also confirmed this differential expression pattern for *MUC12* and *MUC17* in the independent validation data. Previous reports have raised the question about whether the genomic sequences for *MUC11* and *MUC12* are from closely related or perhaps even the same gene [Byrd and Bresalier, 2004]. Correlation analysis of *MUC11* and *MUC12* probesets show a strong, positive correlation at the lower end of the probeset expression range with a weaker correlation as expression increases (data not shown). This correlation profile could be due to increased variability at higher expression levels or, possibly, because the expression levels in the distal colon (where they are higher) reflect a distinct transcriptional control. Differences in mucin glycoprotein characteristics between the proximal and distal gut, including the degree of sulfation, were demonstrated thirty years ago [Filipe and Branfoot, 1976, Bara et al., 1984].

In addition, while previous research has suggested that the secreted, gel-forming mucin *MUC5B* is only weakly expressed in the colon [Byrd and Bresalier, 2004], these results show that probesets reactive to this transcript are expressed at a higher level in the distal colon as for the membrane-bound mucins. These data also support earlier reports that transcripts for the estrogen responsive element known as trefoil factor 1 (*TFF1*, alias: *pS2*) are differentially expressed and elevated in the distal colon [Singh et al., 1998].

Many of the expression patterns reported here for humans have been shown to be similarly patterned in the gastrointestinal tracts of rodent models. However, a number of specific genes previously shown to be differentially expressed along the large intestines of mice and rats were not found to be so expressed by us. Such gene transcript targets, include solute carrier family 4 member 1 (alias *AE1*) [Rajendran et al., 2000], and toll-like receptor 4 [Ortega-Cava et al., 2003]. For *TLR4* no significant difference in expression between proximal and distal human samples was seen by RT-PCR in agreement with the microarray data. Using a

commercially available RT-PCR assay, *SLC4A1* mRNA was not detected in any of the validation set (Appended Figure 4.3). On the other hand, these data are in agreement with earlier studies of expression of aquaporin-8 (*AQP8*), a gene whose expression product is suspected to be involved in water absorption in the normal rat colon [Calamita et al., 2001]. *AQP8* is observed to be significantly expressed at a higher expression level in the proximal human colon compared to the distal tissues ($P < 0.01$), data not shown.)

The family of claudin tight junction proteins may also play a role in maintaining the water barrier integrity in the colon [Jeansonne et al., 2003]. Claudin-8 (*CLDN8*) was shown to exhibit higher expression levels in the distal colorectal tissues and this observation was supported by RT-PCR analysis (see Figure 4.3). Conversely, claudin-15 (*CLDN15*), which is also believed to be localized in the tight junction fibrils was expressed more highly level in the proximal colorectal tissues [Colegio et al., 2002].

6.4.3 The nature of gene expression change along the colon

While one goal of this work was to understand which gene transcripts are differentially expressed along the colon, a second aim was to explore the nature of putative expression changes along the proximal-distal axis in region or segment-specific detail.

Two broad patterns of statistically significant transcript expression change was observed along the colorectum. The major pattern is described by those 65 probesets that were well fitted by a two-segment expression model. The expression of these transcripts appears to be dichotomous in nature - elevated in the proximal segments and decreased in distal segments, or vice-versa.

A second set of 50 probesets do not display a dichotomous change, but rather show a significant improvement in fit by applying the expression data to a five-segment model supporting a more gradual expression gradient moving along the colon from the cecum to the rectum.

These two characteristic expression patterns hint that gene expression along the proximal-distal axis is perhaps coordinated by two underlying systems of organization.

The majority of differentially expressed transcripts in the adult normal tissues measured here are expressed in a pattern that is consistent with a midgut vs. hindgut pattern of embryonic development. Further, multivariate methods including sPCA and canonical variate analysis (data not shown) also suggest that the primary source of variation among these data are explained by the proximal vs. distal divide. In a recent study Glebov et al. found that the number of genes differentially expressed between the ascending and descending colon in the adult is substantially larger than the number of genes likewise identified in 17-24 week old fetal colons. Glebov et al. hypothesize that the gene expression pattern of the adult colon is possibly set concurrently with expression of the adult colonic phenotype at 30 weeks gestation or perhaps even in response to post-natal luminal contents of the gastrointestinal tract. While gene expression in the fetal colon was not explored, patterns of gene expression were observed in the adult that support a proximal-distal expression model consistent with the midgut-hindgut embryonic origins.

Most (41 of 50) of those transcripts that exhibit a gradual expression change between the cecum and rectum exhibit a prototypical pattern of increased expression increasing from the cecum to the rectum. This pattern is not observed in the midgut-hindgut differential transcripts where the number of transcripts elevated proximally is approximately equal to the number elevated in the distal region. I propose that the characteristic distally increasing pattern in those transcripts could be a function of extrinsic factors in comparison to the intrinsically defined midgut-hindgut pattern. Such factors could include the effect of luminal contents that move in a unidirectional manner from the cecum to the rectum and/or the regional changes in microflora along the large intestine. Further work will be required to investigate whether such extrinsic controls are working in a positive manner of inducing transcriptional activity or through a reduced transcriptional silencing.

To explore the expression of genes in concert along the colon, principal component analysis was also applied to these expression data. There is strong evidence for a proximal versus distal gene expression pattern with these multivariate visualization techniques. Though multivariate results do not exclude a subtle proximal-distal gradient, the apparent bimodal nature of the multivariate plots suggests that the major source of expression variation in these tissues is consistent with a midgut- vs. hindgut-derived pattern.

6.5 Conclusions

These data confirm that transcript abundance, and perhaps transcriptional regulation, follows two broad patterns along the proximal-distal axis of the large intestine. The dominant pattern is a dichotomous expression pattern consistent with the midgut-hindgut embryonic origins of the proximal and distal gut. Transcripts that follow this pattern are approximately equally split into those that are elevated distally and those elevated proximally. The second pattern is characterised by a gradual change in transcript levels from the cecum to the rectum, nearly all of which exhibit increasing expression toward the distal tissues. I propose that tissues that exhibit the dichotomous midgut-hindgut patterns are likely to reflect the intrinsic embryonic origins of the large intestine while those that exhibit a gradual change reflect extrinsic factors such as luminal flow and microflora changes. Taken together, these patterns constitute a gene expression map of the large intestine. This is the first such map of an entire human organ.

This understanding of gene expression variation in the normal large intestine provides a strong foundation for the primary aim of this thesis: the analysis of gene expression in neoplastic colorectal tissues.

Chapter 7

Discovery of Gene Expression Markers for Colorectal Neoplasia

7.1 Aim

This chapter describes the discovery of biomarker candidates for colorectal neoplasia derived from two sources of discovery data. These biomarker candidates were used to construct a unique custom oligonucleotide microarray which was then used to test hypotheses generated from these discovery results in independently derived clinical specimens. The results of hypothesis testing of these candidates is discussed in the following chapters.

To improve discovery, two sources of data were analysed 1) quantitative RT-PCR of transcripts discovered by differential display and 2) conventional oligonucleotide microarray data. These discovery data sets are comparatively large. While unpublished, the differential display data used here were generated in one of the earliest studies to focus on colorectal adenomas and normal controls. The microarray data, on the other hand, was the largest set available when purchased in 2004. Importantly, the microarray data also included a large sample (42 tissues) of non-neoplastic disease controls with evidence of colitis.

Both univariate and multivariate techniques were applied to discover diagnostic

expression patterns for testing.

7.2 Differential display discovery

7.2.1 Nucleotide sequences to genes

A team led by Dr. Rob James and Prof. Graeme Young (Flinders Medical Centre) carried out differential display-PCR to comprehensively analyse up-regulated RNA transcripts in a large panel of tubular, tubulovillous, and villous adenomas (See Section 5.2.1, p. 75). Sequential rounds of panning identified a panel of 354 transcript candidates (148 known genes and 206 previously uncharacterised transcripts) to be consistently up-regulated in adenomatous tissue extracts compared to non-neoplastic controls [James and Kazenwadel, 2002, James, 2001]. This differential display discovery research did not explore down-regulated gene expression targets.

To annotate these transcript sequences to putative genes the author designed and developed semi-automated bioinformatics tools (discussed in Section 5.5.3). Details of the annotation and presumed human genomic DNA sources of the candidate transcripts are given in Table D.4, p. 265 of the Appendix.

7.2.2 Preliminary validation: RT-PCR experiments

The expression level of the top 67 biomarkers from the differential display research was next measured in an independent test experiment using quantitative RT-PCR in 71 tissue samples (21 normal, 20 tubular adenoma, 26 tubulovillous adenoma, and 4 villous adenoma).

These data were first explored in terms of total fold up-regulation of each candidate biomarker between adenoma tissues and non-neoplastic controls. To discover which subsets of the candidates correctly discriminate tissue class in a

multidimensional space both logistic regression and a multivariate clustering technique was used.

7.2.3 Univariate analysis

Univariate analysis of the RT-PCR results for the top 30 of 67 primer sets measured across 71 tissues and demonstrating a sensitivity/specificity of 70% or greater is summarized in shown Table 7.1.

Table 7.1: Univariate analysis of RT-PCR data measuring 67 RNA transcript targets in 71 clinical specimens. Only transcript targets with a sensitivity/specificity 70% or greater are shown.

Disc. Clone	*P*-Val (MHT)	D.Val(50)	Fold-Δ	Sens-Spec	95CI
12.2f	0.00	2.85	44.97	92.30	86.1-96.2
11.10e	0.00	2.71	246.02	91.30	84.6-95.5
8.2d	0.00	2.70	49.60	91.10	84.4-95.4
11.5b	0.00	2.53	141.14	89.70	82.6-94.5
4.14b	0.00	2.49	33.10	89.30	82-94.2
5.4a	0.00	2.15	15.18	85.90	77.8-91.7
6.10d	0.00	2.17	5.01	86.10	78-91.9
4.11e	0.00	2.03	8.32	84.50	76.1-90.7
8.7bi	0.00	1.88	296.11	82.60	74-89.2
7.13b	0.00	1.82	21.45	81.80	73-88.6
1.6aii	0.00	1.76	28.78	81.00	72.1-88
12.7c	0.00	1.71	19.73	80.30	71.3-87.4
8.19a	0.00	1.70	11.19	80.20	71.2-87.3
5.13d	0.00	1.69	9.16	80.20	71.1-87.2
6.12a	0.00	1.51	30.29	77.60	68.3-85.2
5.14j	0.00	1.50	16.84	77.40	68.1-85
8.12b	0.00	1.63	4.07	79.30	70.1-86.5
9.2d	0.00	1.50	8.17	77.40	68.1-84.9
3.2c	0.00	1.40	7.37	75.80	66.3-83.7
9.4gclone5	0.00	1.42	3.82	76.10	66.6-83.9
11.10b	0.00	1.30	12.23	74.20	64.5-82.3
2.1c	0.00	1.25	27.49	73.40	63.8-81.6
9.13c3	0.00	1.27	4.42	73.80	64.1-82

7.13dclone4	0.00	1.26	4.78	73.60	63.9-81.7
2.13aclone5	0.00	1.22	5.41	73.00	63.2-81.3
3.12eclone3	0.00	1.15	5.63	71.70	61.9-80.1
11.2d	0.00	1.06	8.09	70.20	60.3-78.7
1.1g	0.00	1.07	5.05	70.40	60.5-79
6.12b	0.00	1.05	10.28	70.00	60.2-78.7
9.8g	0.00	1.08	3.10	70.50	60.6-79.1

Twenty-seven (27/67, 38%) targets were found to exhibit five-fold or greater increased mean expression in adenomas relative to the normal controls. Three targets exhibited a cross-validated sensitivity and specificity by ROC-midpoint analysis $\geq 90\%$.

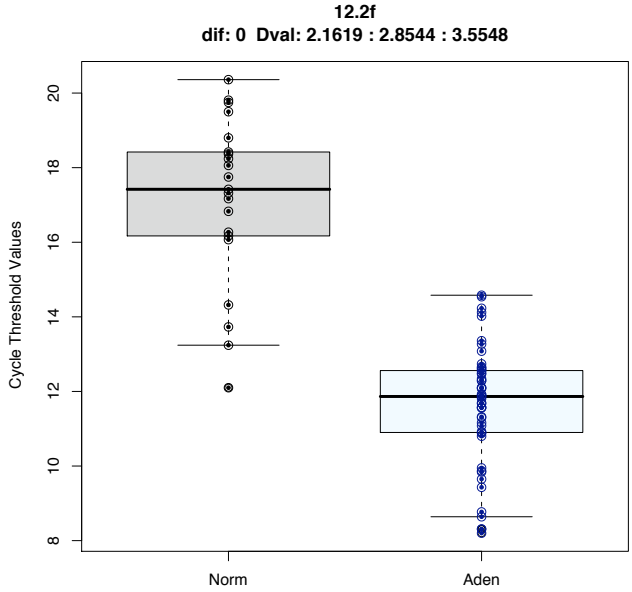


Figure 7.1: RT-PCR measuring using primer set for clone 12.2f demonstrated the highest sensitivity of 92.3%.

The most sensitive univariate transcript was for clone 12.2f with a D-value of 2.85, corresponding to a sensitivity/specificity of 92.3% (95%CI=86.1-92.2%). A boxplot of clone 12.2f is shown in Figure 7.1. While it was not known at the

time of this discovery work, some of these clones were subsequently determined to contain transcripts which are transcribed from separate regions of a common gene locus. In particular, the clone for 8.2d, shown in Table 7.1 as the third most differentially expressed clone, was predicted to correspond to the same gene locus as for clone 12.2f. This relationship is supported by the strong correlation in expression levels observed between these targets shown in Figure 7.2.

Eight of the univariate candidate RNA markers were encouraging as individual biomarkers with observed D-values greater than 2.0, although there was no single target which perfectly separated the data.

7.2.4 Multivariate analysis

Multivariate techniques where then applied to these data to discover multiple marker expression patterns that discriminate adenomas from normal controls.

Logistic regression modeling

Logistic regression models were used to explore class discrimination using multiple gene sets. Given the manageable data set size (71 observations $\times$ 67 genes) an all-subsets algorithm was used to test every possible 2-target and 3-target subset of the 67-target data set. Predicted phenotype for each observation was then compared to the true phenotype values to estimate sensitive, specificity, D value, etc. for each model. As these models were ranked and chosen based on these post-hoc metrics, the performance estimates are (possibly severe) overestimates of the expected performance [Hastie et al., 2001]. Nevertheless, these metrics provide a convenient and objective measure for evaluating and comparing models.

There are 2,211 different combinations of 67 targets tested two-at-a-time. In this model space, there were twenty (0.9%) unique combinations of the PCR targets that perfectly separated the 50 adenomas from the 21 normal tissues. In other words, the two classes were linearly separable in these sub-spaces.

Of the 47,905 unique combinations of three-plex targets, 1,476 (3.0%) yielded a perfect positive predictive value.

K-Nearest Neighbor analysis

In addition to testing the linear separability of phenotypes, k-nearest neighbor clustering was used to explore sample-to-sample relationships. This technique was previously applied to a publicly available colon tumor vs. normal data as described in Li et al. [2001a]. Details concerning our implementation are provided in Section 5.5.11.

A range of k values ($1 \leq k \leq 5$) were evaluated and $k = 3$ was empirically determined to yield manageable results. As expected, increasing k results in a reduced tissue classification as the threshold for unanimous agreement rises while k less than three yielded a higher number of "successful" models.

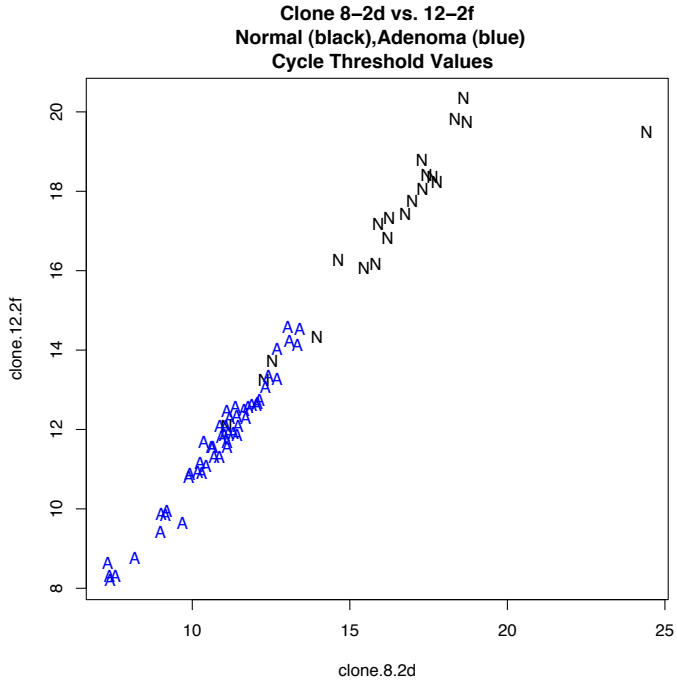
Using a custom application written by the author, all low dimensional ($p \leq 4$) combinations of the candidate expression space were tested. The results for these investigations are shown in Table 7.2. Phenotype classification using the best subsets in two and three dimensions are shown in Figures 7.2 and 7.3, respectively.

Table 7.2: All variable subset analysis of all 2-,3-, and 4-dimensional combinations of RT-PCR validation data using k-nearest neighbor clustering

p-Dimension	Possible Combinations	Results
2	2,211	Best achieved (68/71) Markers: 8/2d with 11/10a; 12/2f&11/10a; 12/2f&3/16b-clone4
3	47,905	Best achieved (70/71) Markers: 8.2d, 4.14b, 4.18e
4	766,480	Best achieved (70/71) Six (6) 4-plex panels.

As the problem space for all possible five-marker combinations for 67 targets is very large, an all variable subsets analysis of combinations of sets greater

Figure 7.2: Best 2-plex shown using RT-PCR ct values for clones 8-2d and 12-2f. There are two interesting points to note from these data. First both clones (i.e. axes) appear to be nearly separating the phenotypes individual. Also, the two clones are strongly correlated for nearly all specimens. Subsequent sequence analysis confirmed that these clones appear to correspond to different transcripts within the same gene locus.



than four was not attempted. Instead, a genetic algorithm was implemented to explore panels of up to 15 candidates to assess the potential for classification performance improvements by using higher dimensional clustering. A similar approach was previously used in Li et al. [2001b] to search the p-dimensional expression landscape for optimum or near optimum marker sets.

Using a genetic algorithm sets of 5-, 8-, 12- and 15-dimensional panels were explored by cluster analysis. Surprisingly, cluster analysis using a unanimous $k = 3$ decision rule for testing combinations of up to 15 transcripts was not able to achieve perfect classification of all 71 normal and adenoma tissues although many near perfect (70/71) (98.6%) solution sets were identified. These near-perfect solutions were not surprising given the demonstration that these results can be achieved with just 3- and 4-member data subsets as described above.

Principal component analysis

A principal component analysis (PCA) (see Section 5.5.13) in the full 67-dimensional data was applied to observe experiment-wide sources of variance. A plot of the 71 observations projected onto a Cartesian system of the first two principal components is shown in Figure 7.4.

Inspection of the first and second principal components of the full data set suggests that the largest source of variance in these data correlates well with the neoplastic state of the tissue. This observation was not surprising as these genes were selected in the first case by means of differential display. Other potentially explanatory co-variates were *not* tested as no further data (e.g. gender, age, site) were available.

Repeated PCA testing in subsets of the 67 targets in these data confirmed that the strongest observed class separation occurred in the subset of just the four (4) most differentially expressed targets as determined by linear modelling (limma). Interestingly, while univariate analysis demonstrated that more than half (58%) of the targets showed statistically significant increased expression

in the adenomas versus the normal tissues ($P \leq 0.05$, Bonf. corrected t test), PCA of the bottom 60/67 (89.5%) targets suggests that the information content of these genes is of marginal utility compared to the top markers. A PCA of these 60 targets (i.e. with the top 7 targets removed) is shown in Figure 7.5. Compared with the PCA shown in 7.4 constructed using all data, Figure 7.5 shows much weaker phenotype class separation.

7.2.5 A closer look at mis-classified specimens

Using KNN clustering, one 3-target and six 4-target marker combinations were able to achieve near perfect class discrimination of 70/71 (98.6%) observations. Interestingly, one particular normal tissue (A7) was misclassified in all but one (6/7) of these experiments with the normal tissue (C5) misclassified in the remaining experiment. An analysis of which tissues are misclassified in all 3- and 4-dimensional clustering experiments is shown in Figure 7.6.

Given the intriguing contribution of two particular normal tissues (A7 and C5) to mis-classification, the data were re-analyzed using the KNN algorithm on a data set with these two observations removed. Analysis of all three-target combinations yields eight unique 3-transcript panels that perfectly classify the remaining 69/69 (100%) tissues.

7.3 Discovery using full genome microarrays.

In addition to the differential display data, genome-wide transcript changes using commercially available oligonucleotide microarray data in 548 colorectal tissues were also explored.

These microarray data provided a number of benefits to this research. First, the full genome gene chips enabled analysis of both down-regulated and up-regulated transcripts. The differential display research did not characterise potentially

down-regulated markers. Also, the large number of tissues provided a much better understanding of expression variability in all phenotypes. In particular, the analysis of 222 non-diseased specimens plus 42 tissues with evidence of colitis provided a relatively large set of non-neoplastic controls to understand normal tissue variation and non-neoplastic disease affects on gene expression change. The analysis of longitudinal expression changes along the colon described in Chapter 6 exemplifies the opportunities presented by these data.

For a complete description of these data see Section 5.2.2.

7.3.1 Quality control

A quality control analysis was performed to remove arrays not meeting essential quality control parameters. A detailed description of quality control methods that were used is described in Appendix B. Briefly, published and novel quality control metrics were used to identify extreme outliers of the data with a reasonable potential to confound these analyses. In addition to on-chip quality metric testing, quality control methods included analysis of RNA quality data provided by GeneLogic, pathology report information for inconsistencies as well as review of selected histological images of source tissues. From an initial database purchase of 548 tissues, 454 were selected for analysis after quality control screening.

7.3.2 Principal components analysis

The full set of $N = 454$ observations by $p = 44,928$ probesets were first explored using principal components analysis. Visual inspection of the data projected onto the first two principal components strongly suggests two distinct sub-populations within the data. By repeatedly overlaying this plot with potential explanatory co-variates data, e.g. gender, age, assay lot numbers, technician details, *etc.*, visual inspection demonstrated that the only tissue descriptors which correlate with these two sub-populations relate to neoplastic status. The

resulting PCA in the full data set annotated by disease state is shown in Figure 7.7. For illustration, PCA plots testing other potential co-variates to explain these sub-populations are Appended as figures in Section D.4.1, p. 271.

Based on this analysis, the primary source of variance through these data relates to the neoplastic state of the tissue. Inspection of Figure 7.7 indicates that both the first and second eigenvectors play a role in neoplastic vs. non-neoplastic class discrimination. In other words, neither the first nor the second eigenvector is able to discriminate the two populations independently. This pattern was also observed in the differential discovery data shown in figure 7.4 which was generated using mRNA transcripts believed to be differentially expressed in colorectal adenomas. The conclusion that neoplasia is the largest source of gene expression variance in the full genome chip data is well supported.

7.3.3 Genes differentially expressed in neoplastic tissues

From 44,928 probesets, more 11,000 probesets were differentially expressed by moderated t-test ($P < 0.05$, Bonferroni correction for multiple hypothesis testing (MHT)). To reduce the number of differentially expressed transcript candidates to a manageable number for subsequent study, an additional absolute fold change filter of 2-fold or higher mean expression change between phenotypes was used. A summary of the results of differential expression analysis is shown in Table 7.3. The 2-fold cutoff, while arbitrarily chosen, is commonly applied in the literature and is practical in relation to the minimum differential expression levels likely to be useful in a subsequent diagnostic assay. With this further selection criteria applied, 108 probesets were expressed higher in neoplastic tissues relative to non-neoplastic controls and 338 probesets expressed lower in neoplastic states relative to normal tissues.

For convenient reporting and discussion, these 446 probesets were annotated using the most recent meta-data and annotation packages available for the microarrays. The 108 over-expressed and 338 under-expressed probesets were mapped to 107 and 327 gene symbols, respectively.

Figure 7.3: Best 3-plex shown using clones 12-2f, 11-e and 11-5b.

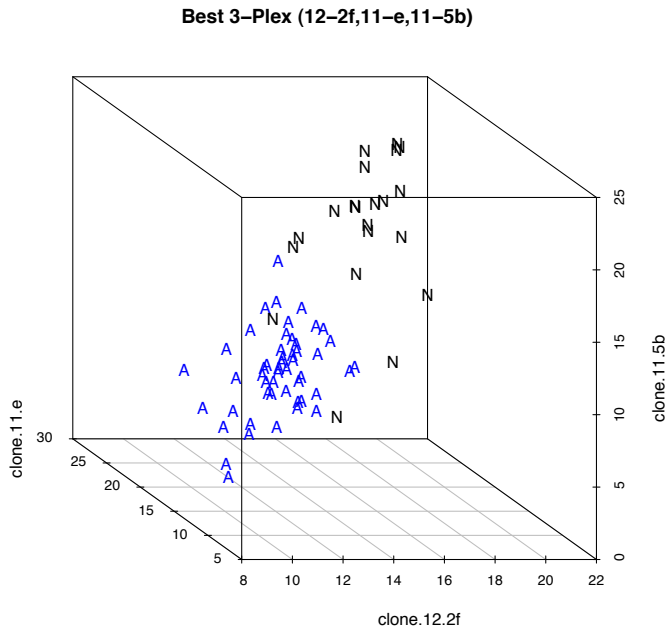


Table 7.3: Analysis of differentially expressed between tissue classes. All P values are Bonferroni corrected for multiple hypothesis testing and fold change is presented in Class B relative to Class A.

Class A	Class B	Diff. $P \leq 0.05$	≥ 2 fold-Δ	Δ down	Δ up
Norm	Adenoma	3161	489	383	106
Norm	Cancer	10897	529	371	158
Adenoma	Cancer	859	181	43	145
Norm	Ad & Ca.	10892	474	356	118
Norm & IBD	Ad & Ca.	11183	446	338	108

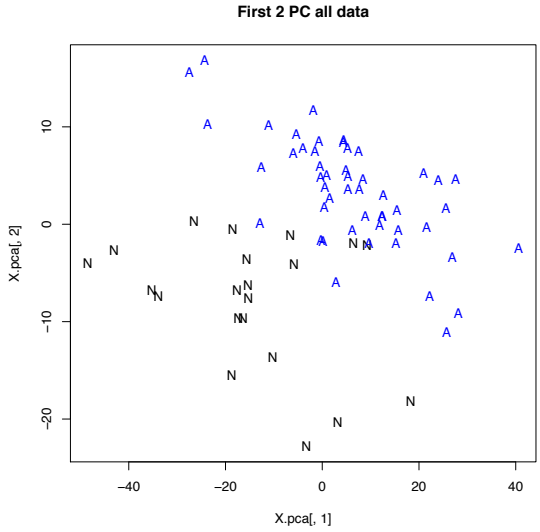


Figure 7.4: Principal components analysis of RT-PCR data collected using biomarker candidates discovered by differential display. Data are identified as either Normal (N) or Adenoma (A). Plot shows evidence of class separation in first two eigenvectors – but interestingly, not in either the first or second eigenvector alone.

Δ-expression	Probeset ID	Probesets mapped to symbol	Not Annotated
UP	108	107	1
DOWN	338	327	11

These up-regulated and down-regulated transcript targets are shown in Appendix Tables D.5, and p. 274 and D.6, p. 276 respectively.

Interestingly, probesets for the *IL8* were found in *both* the up and down gene lists. In fact, there were two probesets corresponding to *IL8* in the "up" list and a different probeset in the "down" list. Careful review of these data reveals that the probeset in the "down" list (but not the "up" probesets) is known to hybridise to more than one Unigene cluster ID which suggests that this probeset is possibly promiscuous for more than one transcript (see naming details discussed in Table 5.3, p.89). This example highlights a general difficulty, and importance, of distinguishing between probeset data and putative gene changes. While microarray gene chip data are often translated to a biological context by annotating probesets to gene symbols, one should bear in mind the nature of the underlying experimental data, i.e. that the data relate to probesets which

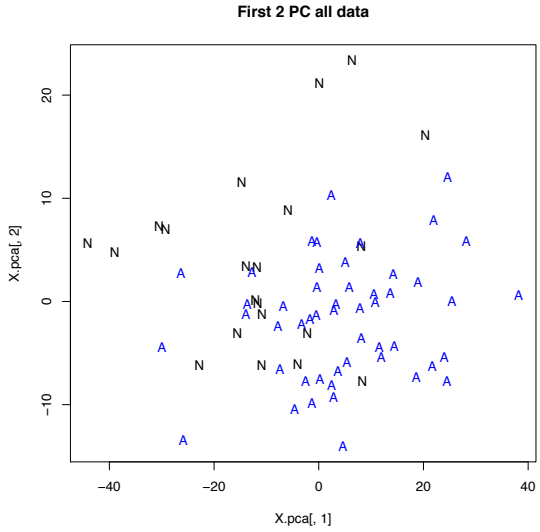


Figure 7.5: PCA plot of the data with the top seven targets removed from the data.

hybridise to oligonucleotide sequences, not to a "gene" per se.

7.3.4 Discovery of neoplasia-specific genes

Finally, differentially expressed patterns of gene expression were analysed to predict transcripts which may be specifically expressed in neoplastic colorectal tissue relative to non-neoplastic controls. In practical terms such probesets should be near background assay concentrations in non-neoplastic tissues with specific, detectable concentrations evident in neoplasia. A transcript that is expressed specifically in neoplastic tissue may potentially encode a translated protein that is also specifically (only) detectable in disease tissue. Ultimately, such a specific protein target measurable in bodily fluids (including e.g. faeces and/or blood) could simplify the design of a diagnostic assay for colorectal neoplasia.

To discover transcripts which are candidates for a qualitative expression pattern, the list of differentially expressed probesets were filtered with a selection criteria aimed at identifying markers specifically expressed in colorectal neoplasia tissues. This filter criteria was based on two simple ideas:

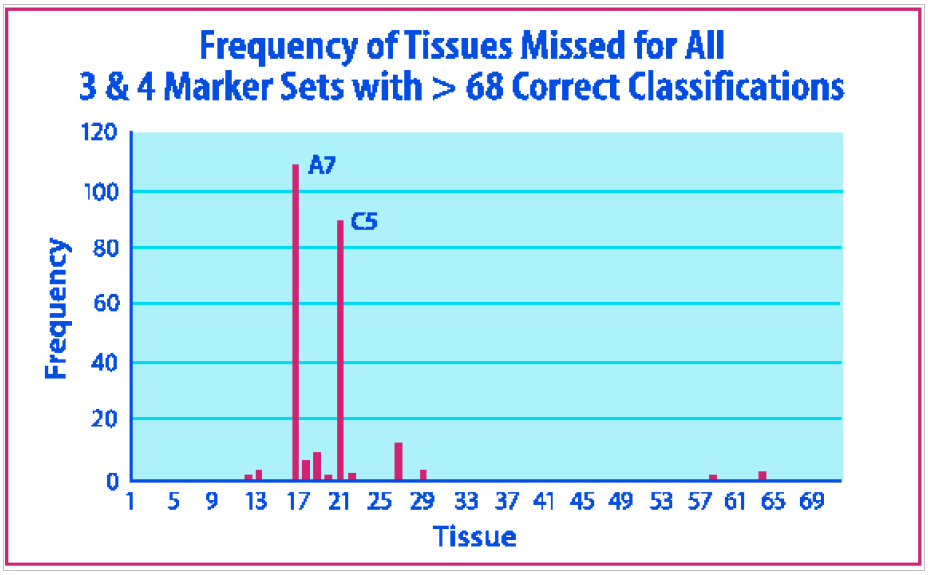


Figure 7.6: Analysis of misclassified tissues using KNN to classify 71 tissue specimens using RT-PCR data. These data suggest that two clinically "normal" specimens (A7 and C5) contribute most often to the set of failed phenotype predictions.

1. That the majority of human transcripts that are present on a genome-wide microarray (e.g. U133) would *not* likely be expressed in the colorectal mucosa; and

2. That microarray binding intensity for such "off" probesets (to labeled cRNA) would reflect technical assay background, i.e. non-specific oligonucleotide binding.

For this analysis, a novel analytical approach detailed in Chapter 5 was developed. To generate a list of hypothetically neoplasia specific (i.e. "turned-on") probesets the non-neoplastic signals were compared with a hypothetical background signal threshold from across all probesets on the microarray. Of course, by design, all probesets in the candidate pool from which the "turned-on" transcripts are chosen were also at least two fold over-expressed in the diseased tissues. Combined, these criteria were used to identify the subset of differentially expressed genes that could be specifically expressed in neoplasia. The expression profile for *NFE2L3*, a representative "turned-on" probeset, is shown in Figure 7.8.

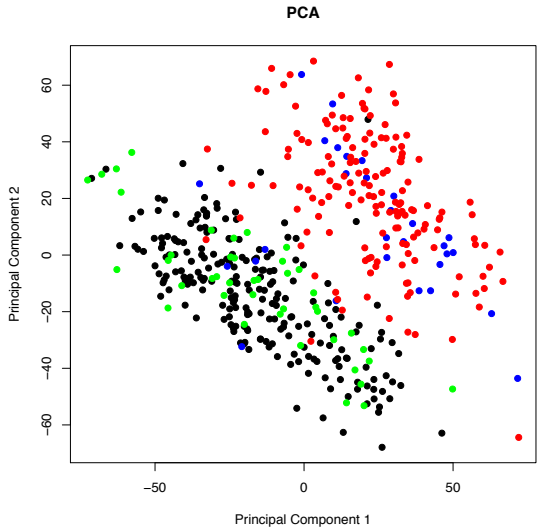


Figure 7.7: PCA of Gene Logic data using all 44,928 probesets. Each tissue is colored by phenotype: Normal (black), IBD (green), Adenoma (blue), Cancer (red). This plot strongly suggests that the data fall into roughly two subpopulations which correlate with neoplastic state.

The subset of 23 transcripts that appear to express a neoplasia specific signal are shown in Table 7.4.

Table 7.4: Probesets/genes which exhibit an expression profile postulated to be "turned-on" in colorectal neoplastic tissues

Probeset ID	Symbol	Description
204702_s_at	NFE2L3	nuclear factor (erythroid-derived 2)-like 3
227140_at	-NA-	-NA-
225806_at	JUB	jub, ajuba homologue (Xenopus laevis)
204259_at	MMP7	matrix metalloproteinase 7 (matrilysin, uterine)
219787_s_at	ECT2	epithelial cell transforming sequence 2 oncogene
238021_s_at	hCG_1815491	hCG1815491
213880_at	LGR5	leu-rich rpt-containing G prot-coupled recptr 5
207850_at	CXCL3	chemokine (C-X-C motif) ligand 3
37892_at	COL11A1	collagen, type XI, alpha 1
222608_s_at	ANLN	anillin, actin binding protein
202286_s_at	TACSTD2	tumor-associated calcium signal transducer 2
241031_at	FAM148A	family with sequence similarity 148, member A
206224_at	CST1	cystatin SN

209309_at	AZGP1	alpha-2-glycoprotein 1, zinc-binding
204475_at	MMP1	matrix metalloproteinase 1 (interstitial collagenase)
202311_s_at	COL1A1	collagen, type I, alpha 1
227174_at	WDR72	WD repeat domain 72
223062_s_at	PSAT1	phosphoserine aminotransferase 1
226237_at	COL8A1	collagen, type VIII, alpha 1
211506_s_at	IL8	interleukin 8
232252_at	DUSP27	dual specificity phosphatase 27 (putative)
204885_s_at	MSLN	mesothelin
214974_x_at	CXCL5	chemokine (C-X-C motif) ligand 5

Conversely, probesets were also identified that appeared to be "turned off" in neoplastic tissues relative to non-neoplastic controls. To identify "turned-off" probesets the filter criteria described above were reversed to find probesets with 1) neoplastic expression levels below our theoretical on/off threshold; and 2) normal signals at least 2-fold higher than disease signals. The expression profile of *ADH1B*, an example probeset hypothesised to be "turned-off" in neoplastic tissues, is shown in Figure 7.9 and a table of all 35 such transcripts is shown in Table 7.5.

It is interesting to note that the "turned-off" expression signal for *ADH1B* observed in the neoplastic tissues exhibits an apparently higher variability compared to the "turned-off" signal observed in normal tissues for *NFE2L3*, an up-regulated gene. One possible explanation for this observation is that neoplastic tissues may contain mRNA contributed from non-neoplastic cells still transcribing this gene resulting in a higher expression variance.

Table 7.5: Probesets/genes which exhibit an expression profile postulated to be "turned-off" in colorectal neoplastic tissues

Probeset ID	**Gene Symbol**	**Description**
204719_at	ABCA8	ATP-binding cass., sub-fam A (ABC1), 8
209613_s_at	ADH1B	alcohol dehydrog. 1B (class I), beta polypep.
230788_at	GCNT2	glucosaminyl (N-acetyl) trnsfrse 2, I-branching enzyme (I blood group)

Continued on Next Page...

Table 7.5 – Continued

Probeset ID	Gene Symbol	Description
228885_at	MAMDC2	MAM domain containing 2
206637_at	P2RY14	purinergic receptor P2Y, G-protein coupled, 14
204931_at	TCF21	transcription factor 21
228504_at	-NA-	-NA-
225575_at	LIFR	leukemia inhibitory factor receptor alpha
231925_at	P2RY1	purinergic receptor P2Y, G-protein coupled, 1
207980_s_at	CITED2	Cbp/p300-interacting transactivator, with Glu/Asp-rich carboxy-terminal domain, 2
227827_at	SORBS2	sorbin and SH3 domain containing 2
209170_s_at	GPM6B	glycoprotein M6B
220376_at	LRRC19	leucine rich repeat containing 19
231773_at	ANGPTL1	angiopoietin-like 1
207080_s_at	PYY	peptide YY
235146_at	-NA-	-NA-
228706_s_at	CLDN23	claudin 23
231120_x_at	PKIB	protein kinase (cAMP-dependent, catalytic) inhibitor beta
202920_at	ANK2	ankyrin 2, neuronal
211549_s_at	HPGD	hydroxyprostaglandin dehydrogenase 15-(NAD)
228854_at	-NA-	-NA-
224412_s_at	TRPM6	transient receptor potential cation channel, subfamily M, member 6
220812_s_at	HHLA2	HERV-H LTR-associating 2
220037_s_at	LYVE1	lymphatic vessel endothelial hyaluronan receptor 1
222717_at	SDPR	serum deprivation response (phosphatidyl-serine binding protein)
205433_at	BCHE	butyrylcholinesterase
203296_s_at	ATP1A2	ATPase, Na+/K+ transporting, alpha 2 (+) polypeptide
219948_x_at	UGT2A3	UDP glucuronosyltransferase 2 family, polypeptide A3
228766_at	CD36	CD36 molecule (thrombospondin receptor)
243278_at	FOXP2	forkhead box P2
203881_s_at	DMD	dystrophin (muscular dystrophy,

Continued on Next Page...

Table 7.5 – Continued

Probeset ID	Gene Symbol	Description
		Duchenne and Becker types)
204940_at	PLN	phospholamban
206664_at	SI	sucrase-isomaltase (alpha-glucosidase)
214598_at	CLDN8	claudin 8
238751_at	SORBS2	sorbin and SH3 domain containing 2

7.3.5 Comparing expression between adenomatous and cancerous tissues

Forty-three probesets were observed to be differentially expressed at least two-fold higher in adenoma tissues relative to cancer tissues and 145 probesets that were expressed at least two-fold higher in cancers relative to adenoma. Lists of probesets up-regulated in adenoma and cancer probesets are included in the Appendix as D.7 and D.8, respectively. Furthermore, several transcripts exhibited expression patterns specific for adenoma and cancer. Examples included *SLITRK6* which demonstrated an adenoma-specific gene expression pattern and *COL11A1* which showed elevated expressions in cancer tissues exclusively. Expression patterns for *SLITRK6* and *COL11A1* are shown in Figures 7.10 and 7.11, respectively.

7.3.6 Multivariate models built from univariate candidates

To explore the benefit of combining differentially expressed candidates, logistic regression models were constructed using the top most differentially expressed biomarkers between selected phenotypes. As expected, a rapid improvement in tissue classification effectiveness was observed by combining gene expression targets. Given the relatively large number of differential targets discovered and

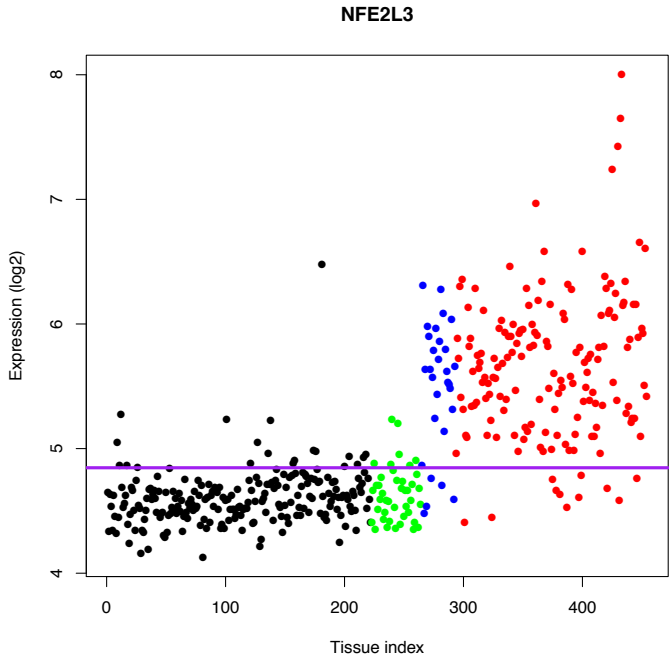


Figure 7.8: Nuclear factor (erythroid-derived-2)-like 2. Tissues are coloured by phenotype: Normal (black), Inflamed (green), Adenoma (Blue), Cancer (Red). This gene exhibits a prototype "turned-on" expression profile in neoplastic tissues including elevated expression as well as tight clustering of non-neoplastic tissue gene expression. For reference a defined "background" cutoff was estimated to be 4.84 (purple line) for this experiment. Note, also, the relatively tight variance observed in the non-neoplastic tissues including the normal and IBD specimens compared to the neoplastic adenomas and cancers.

presented here, only those candidates that exhibit a neoplasia-specific profile as discussed above were used in multi-gene panels. The rationale for this choice was that such candidates, if successfully validated, could simplify future assay development activities.

Starting with the single most differentially expressed neoplasia-specific probeset, consecutive logistic regression models were constructed by iteratively adding probesets one at a time. ROC curves were calculated at each step to compare the classification effectiveness of each iteration. These ROC curves, overlaid against each other, are shown in Figure 7.12 using up to fifteen neoplasia-specific markers. Just ten neoplasia-specific probesets were are able to achieve high

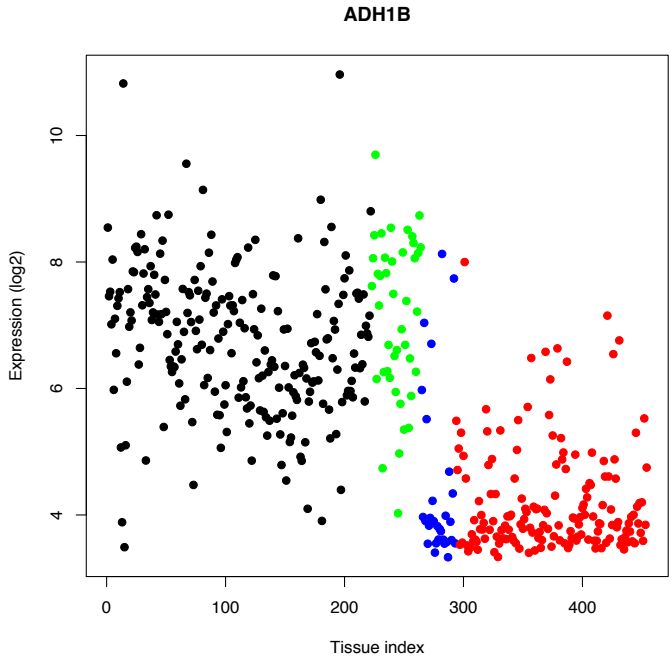


Figure 7.9: Alcohol dehydrogenase IB (class I), beta polypeptide. Tissues are coloured by phenotype: Normal (black), Inflamed (green), Adenoma (Blue), Cancer (Red). This gene exhibits a prototype "turned-off" expression profile in neoplastic tissues. Note the higher variance of the neoplastic "off" tissues relative to the neoplastic tissues which is opposite to the pattern observed for the "turned-on" probesets discussed earlier. One possible explanation for this observation is the contribution of mRNA from non-neoplastic cells in the heterogeneous tumour.

discrimination power corresponding to sensitivity and specificity greater than 97%.

7.4 Pathway analysis by gene set enrichment analysis

Recently Subramanian et al. introduced a new method to analyse large gene expression data sets called "Gene Set Enrichment Analysis" (GSEA) to improve reproducibility and interpretability of gene expression analyses [Subramanian et al., 2005, Bild and Febbo, 2005]. The aim of GSEA is to measure the differ-

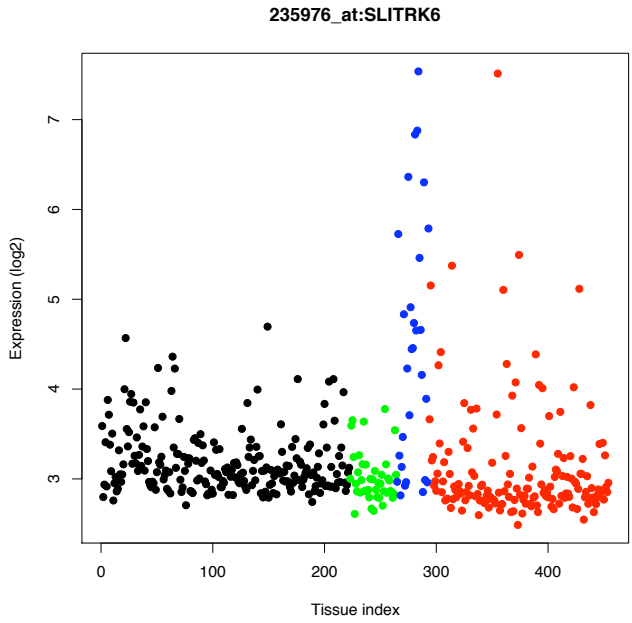


Figure 7.10: SLIT and NTRK-like family, member 6 exhibits elevated expression more frequently in adenomas relative to non-neoplastic and cancer tissues.

ential expression of *a priori* defined subsets of the variables rather than changes in single probesets. In theory, this approach could improve reproducibility and interpretability of microarray analysis by allowing biologists to examine whole gene expression pathway perturbations between phenotypes instead of single genes that are perhaps confounded by noise, etc.. While this assumption was not tested here, gene set analysis was applied to explore the potential for group-wise expression changes in the most well studies pathway related to colorectal neoplasia development, the Wnt pathway.

A priori defined gene sets can be manually assembled or conveniently retrieved from publicly available sources such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [Kanehisa and Goto, 2000, Kanehisa et al., 2008] The introduction of GSEA has led to a number of improvements and/or modifications to the original method including a statistically robust version introduced by Efron and Tibshirani [2006] that is used here.

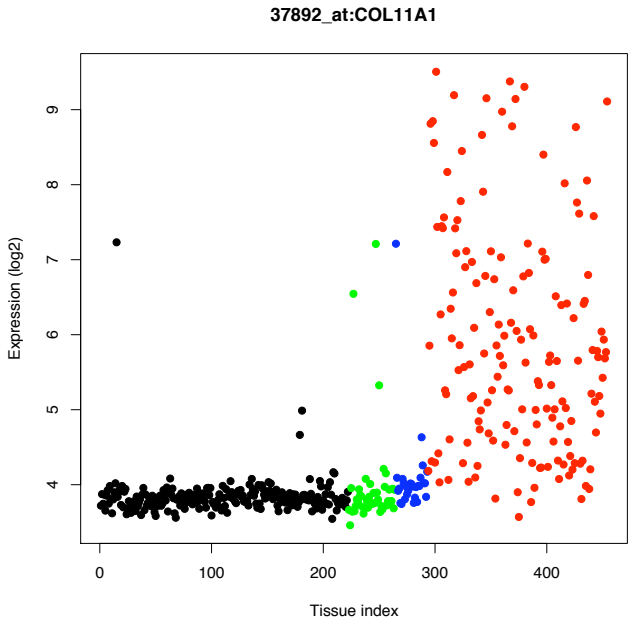


Figure 7.11: Collagen type XI, alpha 1

7.4.1 Wnt pathway analysis

The Wnt pathway is suspected to be perturbed in more than 90% of colorectal neoplasia cases [Giles et al., 2003].

A set of 86 putative Wnt-related gene targets (shown in Appended Table D.1.3) was manually curated from the literature and public domain gene target lists. In particular, most of the targets were taken from Roel Nusse's publicly available curated database available on the Internet (`http://www.stanford.edu/%7ernusse/wntwindow.html`) [Nusse, 2008]. These 86 targets were cross-referenced against the Affymetrix GeneChip annotation to yield 240 probesets[1]. selected to react with transcripts from these genes (hereafter referred to as the EXP-WNT list). For comparison and control, the entire library of 156 KEGG pathway gene sets were also included and tested. The KEGG lists included gene

[1]For convenience we refer to a list of *probesets* as a *geneset*. This is obviously not correct as the biological concept of a geneset in fact refers to a list of gene symbols – not a set of probesets.

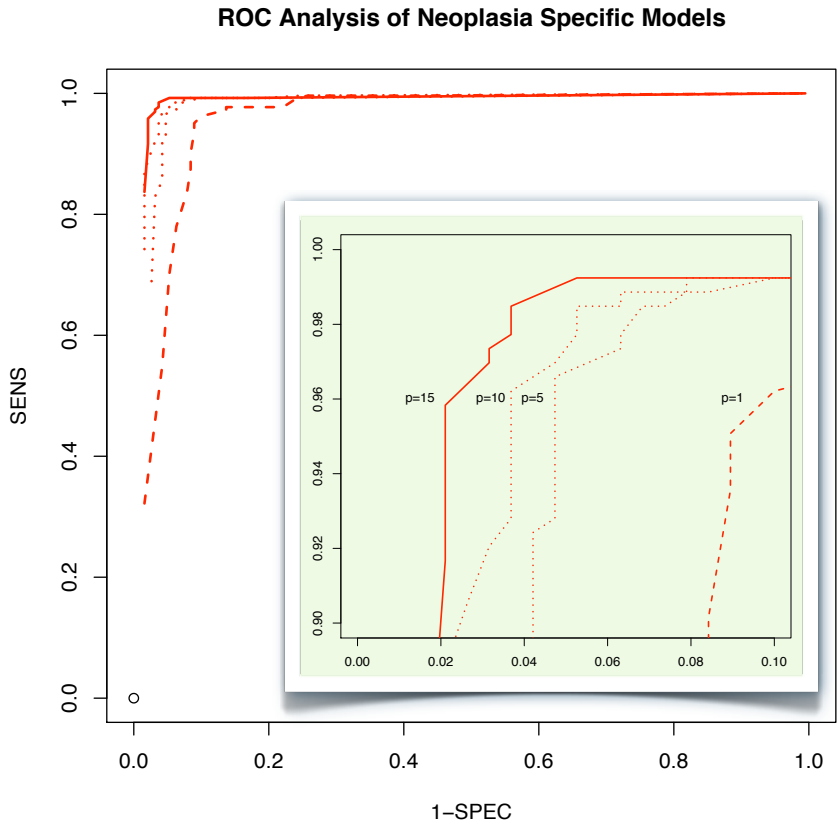


Figure 7.12: ROC analysis of logistic regression modeling with sequentially added probesets, were p is the size of the marker panel. Models of 1, 5, 10 and 15 neoplasia-specific probesets are shown with inset magnified on the region of interest near perfect sensitivity and specificity. A model using 15 probesets shows yields better than 98% sensitivity and 97% specificity.

set lists for Notch, Hedgehog and TGF-β pathways. The complete list of KEGG pathways which were evaluated is shown in D.1.3.

In addition to the manually curated EXP-WNT list, the publicly available KEGG (through BioConductor) list includes a 'Wnt signalling pathway' of 429 probesets. A comparison of the EXP-WNT list with the KEGG list probesets finds that there are only 41 probesets in common. This discordance was not further investigated; however the manually curated list which was predominantly constructed using the results of R. Nusse is well supported based on literature references. One possible explanation for this discordance may involve the degree to which KEGG-based "gene pathways" may involve biochemical networks

which are not strictly related to gene transcription pathways. The EXP-WNT list, on the other hand, was specifically curated to include downstream targets of TCF/LEF1 transcription factors.

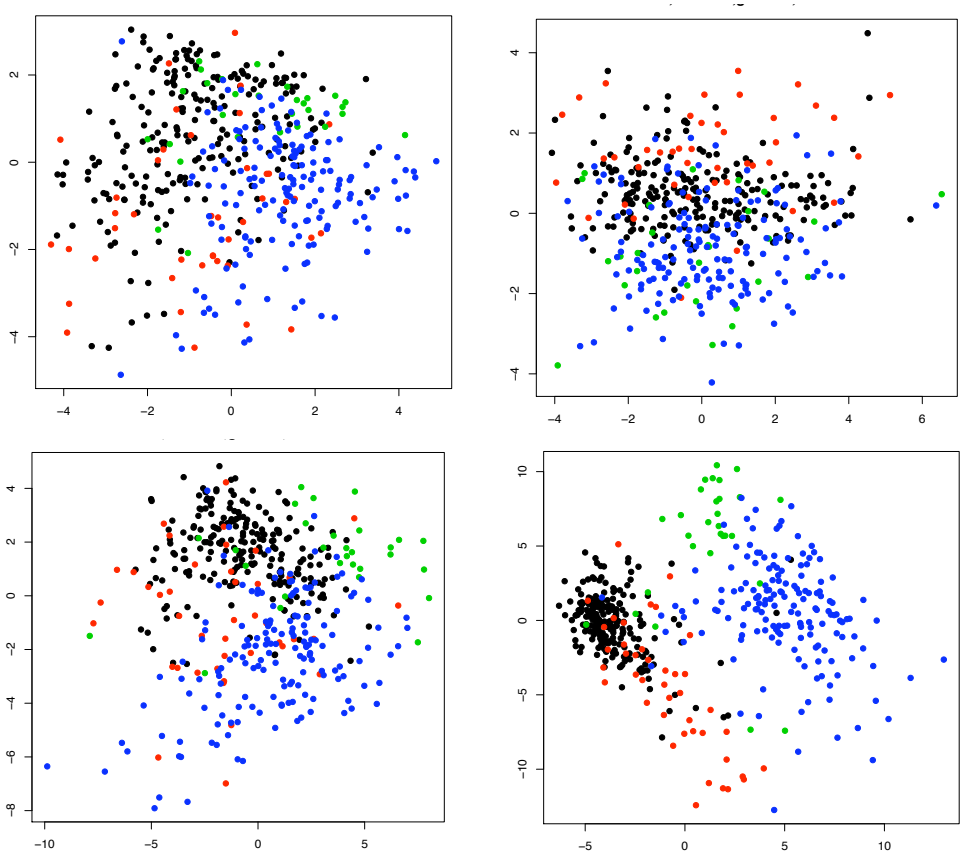
A contrast of the 29 adenoma chips with 222 normal chips by GSA showed that the EXP-WNT list was the ONLY geneset (of 157 independent tests) differentially expressed in adenomas relative to non-diseased controls. None of the gene sets for Notch, Hedgehog or TGF-β nor the KEGG-based Wnt list were shown to be enriched.

In addition to testing the normal vs. adenoma tissues contrasts between i) normal vs. cancer; ii) adenoma vs. cancer; and iii) normal vs. colitis tissues were also explored. Although not related to the primary aims of this thesis, the gene set enrichment analysis for inflamed tissues relative to normal tissues was instructive. Interestingly, this phenotype contrast showed clear evidence of differential expression in inflammatory and infection response pathways. See Appended Table D.4.5, p.287 for detailed results.

7.4.2 Supervised PCA using pathway probesets

In addition to testing known gene set pathways using the GSA algorithm, supervised principal components analyses (sPCA) was applied to the 454 tissue specimens using subsets of the data chosen based on pathway membership. Shown in figure 7.6 are the data projected onto the first two principal components determined using only those genes (probesets) included in the particular pathway. The probesets used for each of the four test pathways (Wnt, Hedgehog, Notch, TGFβ) were the same as used for GSA testing above. Inspection of these sPCA plots confirms that there is strong phenotype clustering within the subspace of Wnt-related probesets. In particular, the sPCA results in the Wnt-related probeset subspace is the *only* experiment carried out in the course of this multi-year research program that separated *all four* phenotypes tested in this research project, i.e. normal, IBD, adenoma, and cancer.

Table 7.6: Supervised principal components analysis exploring the full set of 454 tissues projected onto the first two principal components of subsets of genes based on gene set pathway. For all plots: Black=Normal, Red=Colitis, Green=Adenoma, Blue=Cancer. The four plots show sPCA results for TOP LEFT: Hedgehog pathway (132 probesets); TOP RIGHT: Notch pathway (128 probesets); BOTTOM LEFT: TGF-beta pathway (249 probesets); and BOTTOM RIGHT: Wnt pathway (240 probesets). Note that all probeset lists are based on KEGG pathway annotation except for the Wnt pathway which uses a manually curated list of probesets. See text.



7.5 Literature based discovery

In addition to the biomarkers discovered in the course of these analyses, the scientific literature for hypothetical colorectal neoplasia biomarkers was reviewed (with assistance from collaborators at CSIRO and Flinders Medical Centre). These potential biomarkers were included for representation on the custom microarray design used for subsequent validation experiments.

There is a large body of literature related to gene expression differences between non-neoplastic and neoplastic colon tissues. On the other hand, there are few studies specifically related to adenoma vs. normal or adenoma vs. cancer gene expression. There is also an increasing body of references that include large "lists" of genes suspected to be differentially expressed between normal and cancer tissues. Given concerns about data quality, not all such gene lists were included. The primary quality concerns related to clinical sample handling, data handling (normalisation) and QC scrubbing, statistical methods for discovery, sample numbers, and finally poor interpretation of probeset response related to gene activity. In particular, very few microarray researchers distinguished between probeset binding and possible gene targets (i.e. most researchers use the term "gene" when they should properly refer to "probeset"). Further, where possible, this manually selected list of literature biomarkers was biased toward markers confirmed by RT-PCR.

7.6 Intersection of discovery results

Not surprisingly, there was a significant overlap between the biomarkers discovered between the three discovery methods. Figure 7.13 relates the intersection between each source. On the other hand, the majority of gene symbols identified in each source were, in fact, unique to that source. There are a number of possible reasons for this lack of overlap between these lists. With respect to the "Differential Display" research, the highly novel basis of these experiments, i.e. using a randomly primed transcriptome to sequence differential transcripts in

adenomas in particular, could lead to relatively original findings. The "Microarray" list has a stronger overlap with the literature compared to the "Differential Display" list, perhaps because the commercial nature of the content of the microarrays represents a higher number of "identified" genes. The large number of previously undiscovered differentially expressed genes in the microarrays could be a function of the relatively large size of this experiment. Finally, the inclusion of genes from the literature is more subjective than the other discovery techniques and this list includes, e.g. genes which, while involved in colorectal neoplasia biology, may not be particularly strong targets for gene expression change. For example the *APC* is included in this list of genes for general interest but there is no reason to *a priori* expect that *APC* should be differentially expressed. Finally, both the "Differential Display" and "Microarray" lists represent relatively conservative gene selection lists. An ad-hoc review of a small set of randomly chosen genes from the "Literature" list reveals that some genes *are* differentially expressed in the GeneLogic microarray data, but not to the high degree of significance used to filter selected genes for these experiments.

7.7 Conclusions

This chapter describes the analysis of two unique gene expression data sets that measured concentration of RNA transcripts extracted from multiple colorectal tissue phenotypes. A number of candidate biomarker targets were identified that exhibited differential expression in these discovery data. In particular, candidate biomarkers for colorectal adenomas were identified.

In the differential display data a list of RNA biomarker candidates were found that were elevated in colorectal adenoma tissue extracts relative to normal control tissues extracts. Importantly, this differential display methodology provided the opportunity to discover candidate biomarkers which were not limited to the "established" or annotated transcriptome such as the RefSeq database often used to generate commercial microarrays, including the Affymetrix GeneChips used here. Consequently, a number of the promising candidate targets discovered by

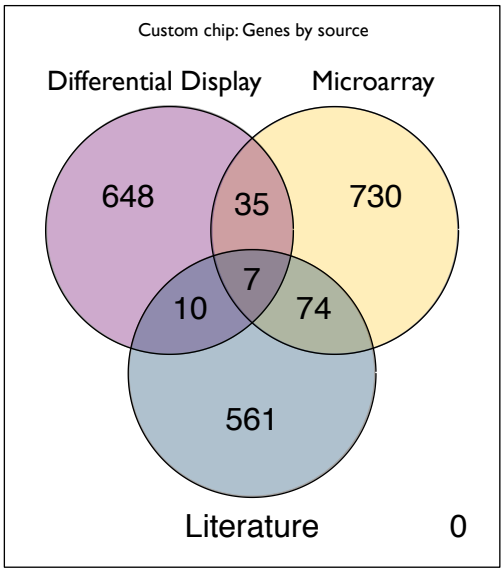


Figure 7.13: Venn diagram illustrating the intersection of gene symbol targets between the three discovery methods: differential display, oligonucleotide microarray, and literature review. Importantly, this intersection is based on gene symbol annotation of both the differential display and microarray data as at the time these analyses were complete in March 2007. Subsequent comparisons (such as shown in Chapter 8) may not precisely align with these exact figures.

differential display were poorly annotated using the publicly available genome bioinformatics tools such as provided by NCBI [NIH/NLM, 2008]. In fact, several of the targets still lack convincing annotation at the time of this report, eight years after the original discovery experiments.

Using a large set of colorectal oligonucleotide microarray data, a second set of biomarker candidates was also shown to be differentially higher or lower expressed in neoplastic tissues relative to non-neoplastic controls. A range of univariate and multivariate statistical tests were applied to these data to reveal probesets which map to human genes which convincingly discriminate colorectal neoplastic tissues from non-neoplastic tissues.

A number of probesets were also observed to discriminate colorectal adenomas from colorectal carcinoma tissues. While the discrimination of phenotypes within the neoplastic phenotype is not a primary goal of this research, the ability to distinguish these tissues based on gene expression patterns may be useful in

some contexts. For example, future research may explore the clinical utility of stratifying adenomatous and cancerous tissues with biomarkers specific for discrete neoplastic stage. Such differentially expressed genes may be informative, for example, about the invasive potential of a given tumour. On the other hand, some of the genes expressed in the cancers but not the adenomas may begin to reflect the more general host response of the body to the growing tumour. For example, a number of collagen genes including (e.g. collagen type XI, alpha 1 *COL11A1*) were shown to exhibit differentially higher expression levels in colorectal cancer tissues relative to adenomas. Such connective tissue genes may be too non-specific with respect to oncogenesis to be useful as cancer biomarkers. Further exploration of this finding is outside the scope of this thesis.

While nearly all previous discovery research employs quantitative metrics of differential expression, a new technique was introduced aimed at filtering the set of candidate biomarkers based on a theoretical on- or off- gene expression pattern in neoplastic tissues. This novel method was motivated by a conceptual bias to discover biomarkers which could enable simplified *in vitro* assays to discriminate neoplastic from non-neoplastic patient samples. This approach presumes that "turned-ON" gene transcripts might lead to a qualitative change in translated protein products downstream.

Interestingly, a large number of microarray probesets (approx. 25%) manifest strong neoplasia vs. non-neoplasia class discrimination in simple univariate comparisons between phenotypes. The relatively large number of univariate targets that were discovered has obviated the need to employ more sophisticated multivariate methodologies to yield a surplus of candidate biomarkers for validation testing. Further, extremely high classification and discriminatory power was be achieved in these discovery data by simply combining strong univariate targets in a multivariate analysis. The strength of this approach is exemplified in Figure 7.12.

The relatively large microarray data-set was also used to test specific hypotheses involving the potential for differential expression between tissue phenotypes in the four major gene expression pathways of the large intestine, Wnt, Hedgehog,

Notch, and TGF-β. From these experiments probesets corresponding to the Wnt pathway genes were identified to be differentially higher in the adenoma (and cancer) tissues relative to the non-neoplastic controls. Group-wise probeset differences were not detected for the other pathways between neoplastic tissues and controls. On the other hand, a high number of inflammatory response pathways were increased in colitis tissue data compared with normal tissue data, consistent with the general understanding of that pathology.

Further evidence of the importance of the Wnt pathway in these data was found using supervised PCA in the Wnt-related probeset subspace. This analysis demonstrated, for the first time, gene expression-based class separation for *all* tissue phenotypes tested here, including normal, colitis, adenoma and cancer. The correlation of neoplasia vs. non-neoplasia with the first two principal components using the Wnt subspace is not surprising given the evidence for an Wnt-related (APC, β-catenin, TCF/Lef, etc.) etiology for colorectal neoplasia. Less well understood is why a Wnt-based sPCA should demonstrate phenotype clustering of the IBD tissues relative to the four other tested phenotypes. This observation was not further explored.

The principal aim of this discovery work in this Chapter was to inform the design of a custom microarray for initial hypothesis testing of these targets. Few of the candidate markers discovered in the previous literature have survived validation [Tinker et al., 2006]. Several of the suggested reasons for this poor validation rate are [Ransohoff, 2004b, Pepe et al., 2001]:

- Many studies are limited by a small data set for discovery;
- There is insufficient attention given to understanding the full range of expression in the non-disease phenotype; and
- There is a lack of good "other" disease phenotype controls.

The data analysed here overcome each of these potential problems. The results of these analysis thus form the core candidates for inclusion in a custom microarray for colorectal neoplasia discrimination which was designed and commissioned by the author. Chapter 8 reviews the first set of independent data collected in fresh

frozen colorectal tissues to test the putative marker hypotheses generated from these discovery data.

Chapter 8

Assessing candidate markers for colorectal neoplasia

8.1 Aims

The previous chapter describes the discovery of RNA biomarkers for colorectal neoplasia in two sources of gene expression data. These markers were combined with biomarkers collected from the literature to construct a custom oligonucleotide microarray. To test the hypothesis that each of these discovered biomarkers is differentially expressed in colorectal tissue, RNA was extracted from 68 independently derived colorectal specimens and assayed using these custom microarrays. This chapter reports the results of these hypotheses testing or "validation" experiments. By demonstrating (or failing to demonstrate) that candidate biomarkers are differentially expressed in an independent set of clinical specimens these experiments address the central question of this research: that gene expression changes can be used to accurately discriminate colorectal neoplastic tissues (both adenomas and cancers) from non-neoplastic controls. For convenience the validation of candidate biomarkers is reported based on the source of discovery data.

8.2 Custom chip design results

A custom microarray for the candidate biomarkers was designed in partnership with Affymetrix, and an initial lot of 90 microarrays was fabricated. Details of the design and content of the custom oligonucleotide microarray platform are given in Section 5.3, p. 78.

8.2.1 Composition of the custom microarray

Probesets were designed for hybridisation against RNA transcripts including well described gene transcripts as predicted by RefSeq and also proprietary RNA transcripts discovered by differential display and sequencing experiments. Probesets from commercially available Affymetrix human genome products were also included comprising both traditional 3' biased probesets as well as probesets from the new human exon microarrays designed to hybridise to target gene exons across the open reading frame of candidate genes.

To select probes against target genes provided by the commercial Affymetrix exon arrays, candidate gene symbols were matched against "transcript cluster ID" according to GenBank records. Corresponding exon probes which hybridise to target transcript cluster ids (usually exons based on RefSeq annotation) were selected for inclusion on the custom microarray. In some cases the number of possible exon probes for a given transcript cluster ID exceeded the available space of the microarray design. For these cluster IDs, a representative selection of probes approximately evenly distributed across the locus was included.

After fabrication and delivery of these custom chips, new annotation routines were written by the author to map each final probeset of the custom microarray to a target gene symbol (including undescribed "LOC" symbols, etc.) based on the currently available map of transcript cluster ID to gene symbol. It should be understood that, for some probesets, this reverse mapping from transcript cluster ID yielded multiple putative gene symbols, i.e. there was a one-to-many relationship of probeset to gene symbols. To avoid bias of gene annotation, all

symbols were mapped individually back to each such probe. One byproduct of this approach is that the final set of "genes" discovered from each source (i.e. differential display, GeneLogic data, and literature) was larger than the original discovery results.

A final disposition of the custom microarray content by probeset type is shown in Table 8.1 and by design source in Table 8.2 . A comparison of the genes

Table 8.1: Analysis of probesets used to fabricate the custom microarrays used for validation experiments.

Probeset type	Probesets
3' biased (U133)	4881
Exon (HuGene)	40083
Custom Diff Display	442
AFFX control	62
Calibration	881
Other custom	133
TOTAL	46482

Table 8.2: Analysis of probesets used to fabricate the custom microarrays used for validation experiments by discovery source.

Source	Probesets	Gene symbols
Differential display	8470	534
Microarray (GeneLogic)	21894	1169
Literature	15114	795

in common which were mapped back to each source of probeset content from the discovery and literature review are shown in Figure 8.1

8.3 Clinical specimens

Total RNA was extracted from 68 fresh frozen colorectal mucosa specimens procured from a tissue bank of colorectal mucosa at Flinders Medical Centre according to a Flinders Medical Centre Ethics Committee approved protocol. The 68 samples include:

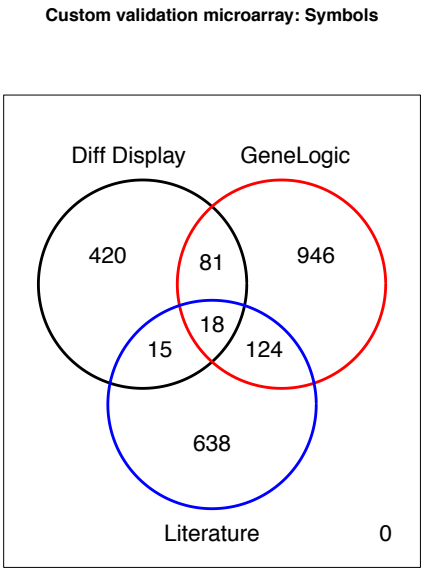


Figure 8.1: A Venn diagram of gene symbols included on the custom validation microarrays annotated using May 2008 libraries. Due to annotation changes in time between the analysis of these validation experiments and the earlier discovery/design, the gene symbol overlap will not match the Venn diagram pattern shown in Figure.7.13, p.150. The inclusion of HuGene probesets also introduces a potential for more gene targets because some exon probesets hybridise to transcript cluster IDs corresponding to more than one gene symbols.

- 30 samples extracted from histologically "normal" colorectal mucosa;
- 19 samples from adenomatous tissue; and
- 19 samples from adenocarcinoma tissue.

The tissues were not matched to patient. Details of the extraction and RNA purification from these tissues are described in Section 5.4 on p. 81. All RNA extracts were assessed for purity and condition using gel electrophoresis. Only RNA extracts meeting strict quality standards were considered or used for gene chip analysis. Tissue specimens were selected to avoid bias from gender, age, and colorectal location. A description of these 68 tissues is shown in Table 8.3.

The 68 RNA extracts were assayed on the custom microarray using a random hexamer-based DNA labelling procedure. Further details of assay procedures are discussed in Section 5.4.3, p. 84. The use of this method was important as some transcripts (e.g. those discovered by differential display) were not necessarily

Table 8.3: Analysis of tissues used in the test/validation research.

		Normal	Neoplasia
Gender	Female	16	20
	Male	14	18
Anatomy	Proximal	14	18
	Distal	16	20
Age	under 40	1	3
	40-49	0	0
	50-59	4	3
	60-69	8	10
	70-79	13	15
	over 79	4	7
Neoplasia			
	Adenoma		19
	TA		1
	TVA		8
	VA		2
	FAP		2
	Unk		6
	Cancer		19
	Dukes' A		17
	Dukes' B		2

identifiable using a poly-dT-based primer system.

8.4 Quality control analysis of the custom microarray data

Significant attention was given to RNA integrity and tissue processing at all stages of RNA extraction and handling and RNA quality was assessed both immediately after extraction and during specific steps of gene chip processing (see Section 5.4.3, p. 84 for further details). Principal component analysis was used to uncover potential sources of experiment-wide variation and bias.

A total of 110,224 probes corresponding to 46,482 probesets were measured for 68 RNA extracts. A PCA plot using all probeset data (shown in Figure 8.2) provides strong evidence of two experiment-wide sub-populations separated

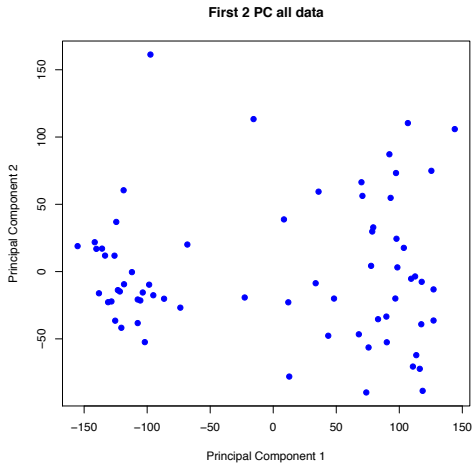


Figure 8.2: Principal components analysis of the custom gene chip data used to test hypotheses related to differential gene expression. There are two populations evident along the first principal component.

along the first principal component. After cross-referencing this PCA plot using all available co-variate data (e.g. age, gender, RNA extraction concentration, etc.) the only tissue descriptor which correlated with these two sub-populations was neoplastic state of the tissue. For comparison, Figure 8.3 illustrates PCA plots highlighted by both age and neoplastic state. These data suggested that, similarly as for the microarray discovery data, the largest source of variance in these data depends on whether a tissue is neoplastic or not.

In Figure 8.4 the 68 tissues are again projected onto the first two principal components, however in this plot the neoplastic tissues are further identified by specific phenotype, either adenoma or cancer. Inspection of this plot suggested that there was an adenoma versus cancer variance correlation in the second principal component. *This normal-adenoma-cancer phenotype separation has not been previously observed in the reported literature although there was a suggestion of neoplasia phenotype discrimination in the supervised PCA plot using Wnt-related genes discussed in Section 7.4.2 on p. 146.* Interestingly, tissue (TB_152_00), which is described as macroscopically and microscopically "normal", clustered with the neoplastic tissues in this PCA plot. After further inquiry with the clinical tissue bank, there was no reason to suspect a tissue

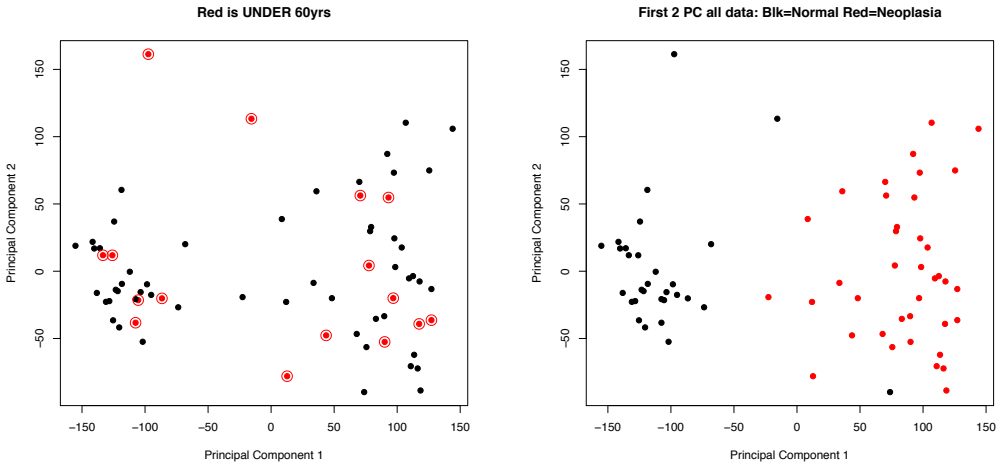


Figure 8.3: Two PCA plots of the full chip data highlighted by age (LEFT: under and over 60 years of age) and neoplastic state (RIGHT: neoplasia vs. non-neoplasia)

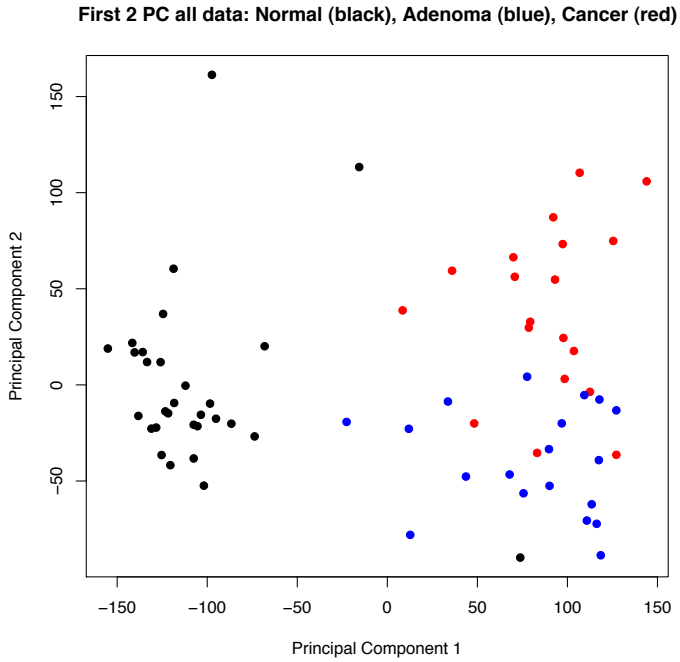


Figure 8.4: PCA plot of the custom chip data highlighted by neoplastic state. This plot provides evidence that gene expression data can be used to discriminate three discrete phenotypes related to colorectal neoplasia: non-neoplasia normals versus adenomatous tissues versus cancerous tissues.

handling or annotation error. It is interesting to note, however, that this tissue also expressed microRNA profiles which are also typical of neoplastic [Michael, 2008].

8.5 Hypothesis testing of differential display candidates

8.5.1 Custom probes against sequence IDs

Differential display research resulted in the identification of 328 nucleotide sequences [James, 2001] which were differentially over-expressed in adenoma tissues. Preliminary validation measured RNA concentration of 67 of these 328 candidates by RT-PCR using sequence specific primers in 50 neoplastic tissues and 21 non-neoplastic controls (See Section 7.2, p. 123). These preliminary experiments provided the first validation evidence that the differential display discovered genes are over-expressed in adenoma tissues.

To further test hypotheses that these candidate targets are differentially expressed in cancerous and adenomatous tissues, probesets designed to hybridise to these 328 nucleotide sequences were measured on the custom microarray. After correcting for redundant sequences, the custom microarray included probeset targets against 304 raw sequences corresponding to 397 unique probesets. For convenience, these candidate biomarkers are referred to herein by their unique "Sequence ID" (SeqID) description. A number of these SeqIDs align with GenBank records with little or no gene annotation detail or available description.

Of these 304 raw Sequence IDs, 172 targets (57%) had a significant mean expression level higher in the neoplastic tissues (i.e adenoma and cancers) relative to the non-neoplastic controls ($P \leq 0.05$ with Bonferroni multiple hypothesis test (MHT) correction).

Eleven sequence targets demonstrated a sensitivity and specificity greater than

90%, shown in Table 8.4 based on these custom probesets. The highest univariate sensitivity and specificity was 96.9% for Sequence ID 302 (D value of 3.86). A plot of comparative expression levels for SeqID 302, which is believed to hybridise to transcripts from the gene *S100A11*, is shown below in Figure 8.5.

Table 8.4: SeqIDS with a univariate sensitivity & specificity $\geq 90\%$ for neoplasia.

SeqID	D val	Symbol	Fold-Δ	Sens=Spec
302	3.74	S100A11:LOC730558:...	3.86	96.9
66	3.15	SLC12A2	3.1	94.2
309	2.95	SLC12A2	3.19	93
296	2.79	APEX1	1.84	91.8
9	2.75	LOC731404:LOC729194:MYC	2.76	91.5
336	2.74	-NA-	2.98	91.5
62	2.75	S100P	6.9	91.5
20	2.69	-NA-	3.4	91.1
119	2.64	CCDC130:C19orf53	2.15	90.7
102	2.63	GALNT6:ELA1	3.36	90.6
263	2.63	NA:CG_63_Seq_ID263_st	1.99	90.6

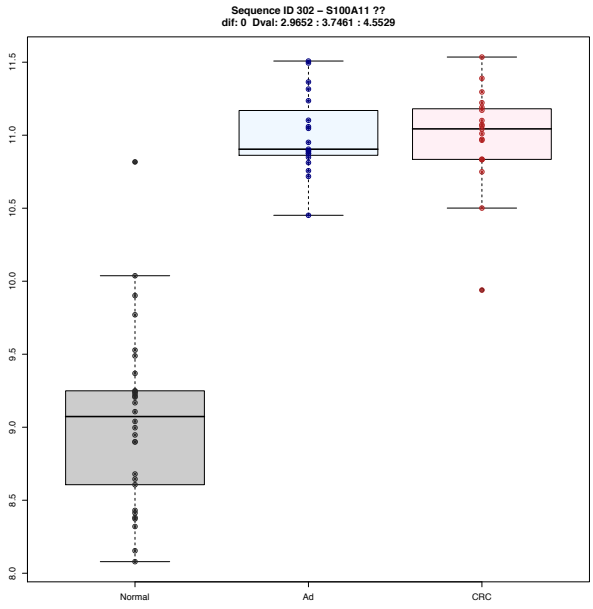


Figure 8.5: Sequence ID 302, putative BLAST annotation: S100A11

A complete list of validated differential display targets using these custom probesets for neoplasia and adenomas is provided in Appended Tables D.10, p. 287

and D.11, p. 290 respectively.

8.5.2 Commercial probes for presumed gene symbols

In addition to the custom probesets designed to (specifically) hybridise to proprietary (and potentially novel) target transcript sequences, the custom microarray also included commercial Affymetrix probesets that target the *presumed* gene expression transcripts corresponding to the 328 sequences. Putative gene symbols corresponding to nucleotide sequences were determined using BLASTn and in-house software to predict a likely molecular identity for each candidate (See Section 5.5.3, p.87). A list of SeqIDs and putative gene symbols is shown in Appendix Table D.4, p. 265.

Of the 328 patent candidates, 289 (88%) aligned with high sequence similarity to a "known" gene or transcript cluster ID for which commercially derived probesets were available. 197/289 (68%) such biomarkers showed a mean transcript expression level in corresponding commercial probesets that was statistically elevated in the neoplasia tissues ($P \leq 0.05$ with Bonferroni MHT correction).

Importantly, 21 of the candidate sequences were differentially expressed in both commercial and custom probeset content by at least a 2-fold change. These confirmed biomarker candidates are shown in Appended Table D.12, p. 293.

8.5.3 Multivariate analysis: logistic regression

Many of the target nucleotide sequences described above demonstrated encouraging class-separation (e.g. neoplasia vs. non-neoplasia) by univariate analysis. Nevertheless, no individual marker yielded perfect class separation on its own for the 68 tissues.

By testing all possible 2-gene combinations of 397 probesets (i.e. probesets against the "raw" sequence) ($N = 78,606$) using logistic regression models, 118 unique 2-gene models were identified that perfectly separated neoplastic from

non-neoplastic tissues. Frequency analysis showed that these 118 duplex sets consisted of 89 unique SeqIDs and that 76/118 (64%) of the subsets included SeqID9, which was identified to correspond to the nucleotide sequence of the *MYC* gene by BLASTn.

Inspection of the univariate response shown by SeqID9 confirms that this target demonstrated high discrimination power for most of the tissue samples. The expression profile of SeqID9 (*MYC*) across all three phenotypes is shown in Figure 8.6.

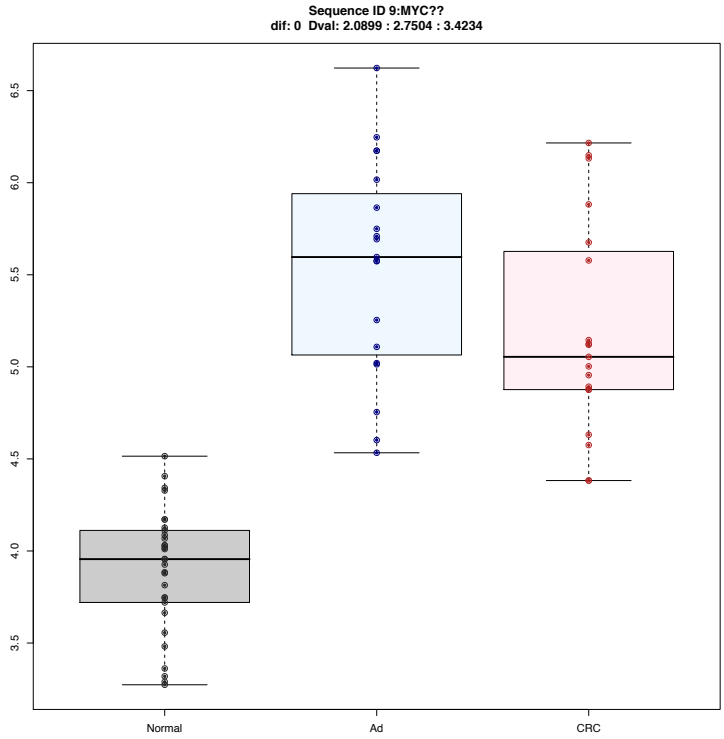


Figure 8.6: SeqID9 (*MYC*. Boxplot shows custom microarray validation experiments in 68 specimens: 30 non-neoplastic controls (Normals), 19 adenomas (Ad), and 19 cancer specimens (CRC). This probeset show approximately 91% sensitivity and specificity for neoplastic tissues in aggregate (Ad+CRC).

8.6 Hypothesis testing of microarray-derived candidates

Results from hypothesis testing of probesets discovered by microarray analysis are described next.

8.6.1 Testing proximal vs. distal expression patterns

In addition to candidate biomarkers of neoplasia, the custom gene chip included probesets that were found in to be differentially expressed between the proximal and distal large intestine. As discussed in Chapter 6, these probesets were validated at the time of discovery in 19 independent tissues. For completeness, these location-specific probesets were also evaluated in these new clinical specimens using the custom chip. In order to avoid possible confounding effects on gene expression introduced by disease, this analysis was carried out using only the 30 non-neoplastic RNA extracts which included 14 samples of proximal origin and 16 samples from the distal colon (See Table 8.3, p.158).

52 (25%) of the 206 probesets previously shown to be differentially expressed between the proximal and distal colon were likewise differentially expressed in the new 30 specimens ($P <= 0.05$,MHT=Benjamini-Hochberg (BH)). Supervised PCA in both the 206 and 52 probeset subspaces again suggested a proximal versus distal pattern although the proximal-distal clustering is not as pronounced as was observed in the original microarray data. The 52-probeset supervised PCA plot is shown in Figure 8.7.

Interestingly, of the 52 probesets found to be differentially expressed in these data, 44 probesets were elevated in the proximal tissues compared to just 8 probesets elevated distally. Thus 44/116 (38%) of the proximal elevations were confirmed while just 8/90 (9%) of the distal elevations were confirmed. This phenomenon was not further investigated.

Finally, the rank of expression change in the list of 206 probesets identified in

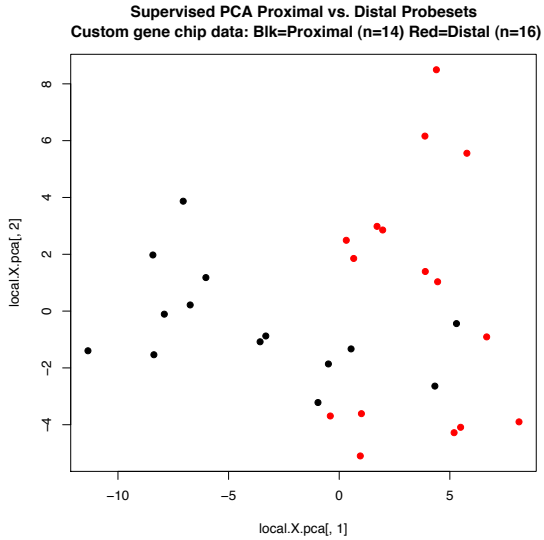


Figure 8.7: Supervised PCA in 30 normal tissues by 52 probesets selected for proximal-distal expression changes.)

the original discovery data was analysed to understand whether the most differentially expressed probesets in the discovery data (i.e. relative rank in original discovery data) were more likely to be confirmed in the new validation data. Figure 8.8 is a histogram of the original rank order (from 1 to 206) by P value for mean difference in the discovery data for the 52 probesets confirmed to be differentially expressed in the test data. By inspection, there are, in fact, more low order (i.e. lower P probesets discovery) probesets that were differentially expressed than high order probesets. Thus, a weak conclusion can be drawn that a probeset that was more differentially expressed in the discovery data was more likely to be confirmed by hypothesis testing in the validation data.

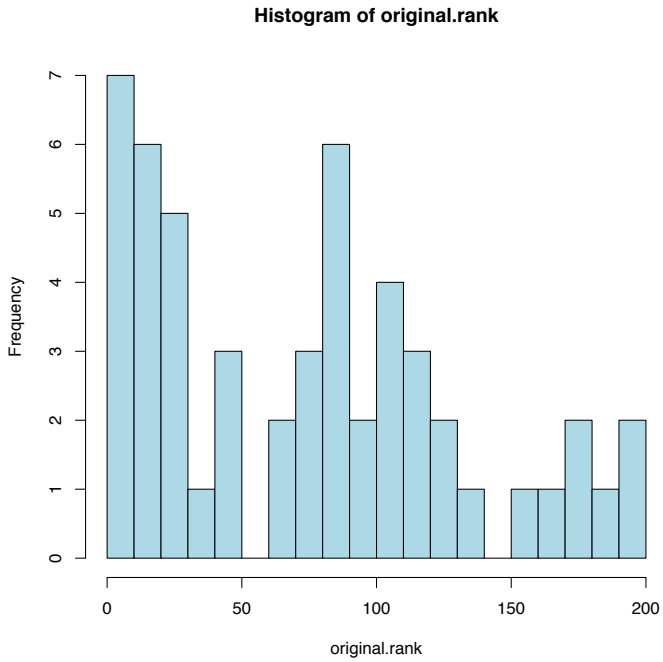


Figure 8.8: Histogram of the original rank from the discovery data of the 52 differentially expressed probesets tested. There is a bias toward lower order rank probesets in these 52 tests. This suggests that the more differentially expressed (i.e the lower the P value) a probeset in the discovery data, the more likely that probeset was to be confirmed in these hypotheses testing experiments.

8.6.2 Hypothesis testing of probesets for neoplasia discrimination

Hypotheses for differentially expressed genes from each discovery phenotype contrast (normal versus adenoma, normal versus cancer, and adenoma versus cancer) were tested according to the same phenotype contrast in the validation data. For convenience a review of observation set sizes is shown below in Table 8.5. Each of the candidate probesets was represented on the custom microar-

Table 8.5: Review of tissue numbers for hypothesis discovery and hypothesis testing data sets.

Contrast	Discovery Data	Validation Data
Normal vs. Adenoma	Normal (161) & Colitis (42) versus Adenoma (29)	Normal (30) versus Adenoma (19)
Normal vs. Cancer	Normal (161) & Colitis (42) versus Cancer (161)	Normal (30) versus Cancer (19)
Adenoma vs. Cancer	Adenoma (29) versus Cancer (161)	Adenoma (19) versus Cancer (19)

rays with two types of content: 1) Probesets identical to the standard HG U133 content as discovered in the original microarray data; and 2) Probes from the HuGene ST1.0 exon array which are designed to hybridise to Transcript Cluster IDs corresponding to gene symbols which mapped from the original HG U133 probeset IDs.

489 and 529 probesets previously shown to be differentially expressed in adenoma and cancer tissues, respectively, were tested. 387 (79%) of the adenoma probesets and 440 (83%) of the cancer probesets were confirmed to be differentially expressed ($P <= 0.05$, MHT=BH) in these test data. In particular, of the 106 probesets shown to be expressed higher in adenomas relative to non-neoplasia in the discovery data, 103 (97%) were likewise determined to be differentially expressed in the test data. An overview of probeset results is shown in Table 8.6. For each contrast, the "HuGene" probes which are designed to hybridise to

transcript cluster IDs identified by the standard HG U133 discovery data were also tested. Results for these exon-level probes are also included in 8.6. For

Table 8.6: Review of probeset numbers for hypothesis discovery and hypothesis testing data sets. Note that an "up" probeset means a probeset differentially higher in the second phenotype relative to the first phenotype.

Contrast	Discov	UP	DOWN	Valid	UP	DOWN
Nrm vs Ad	489	106	383	387 79%	103 97%	284 74%
HuGene	10052	2239	7813	7044 70%	2117 95%	4927 63%
Nrm vs Ca	529	158	371	440 83%	134 85%	306 82%
HuGene	10025	3139	6886	7859 78%	3069 98%	4740 69%
Ad vs Ca	188	145	43	83 44%	58 40%	25 58%
HuGene	3497	2638	859	1841 53%	1506 57%	335 39%

both adenoma- and cancer-based differential discovery probesets, a high percentage of probeset hypotheses tests were validated. The number of confirmed "up" probesets was higher than the number of confirmed "down" probesets in each set of tests, and the difference between the numbers of confirmed tests was significant by a wide margin ($P < 0.01$, 95%CI for diff: 9-34%). There are no reports in the literature of validation differences between up- and down-regulated transcripts, but it is possible that down-regulated gene targets may be more easily confounded by "contamination" due to the presence of non-neoplastic cells in neoplastic validation tissues. This phenomenon was not further investigated.

The most differentially expressed probeset (based on discovery probesets) for adenoma and cancer was the same for each contrast: PS:280037-HuGene_st from the extended HuGene pool targeted against a Transcript Cluster ID from *CDH3*, the placental form of cadherin 3, type 1. The expression profile of this probesets in this validation data is shown in Figure 8.9. There were 195 HG U133 probesets in common between validated probesets for adenoma and cancer, corresponding to 153 gene symbols. A complete list of overlapping, confirmed gene symbols for colorectal neoplasia, i.e. adenoma and cancer, is shown in Appendix Table D.13.

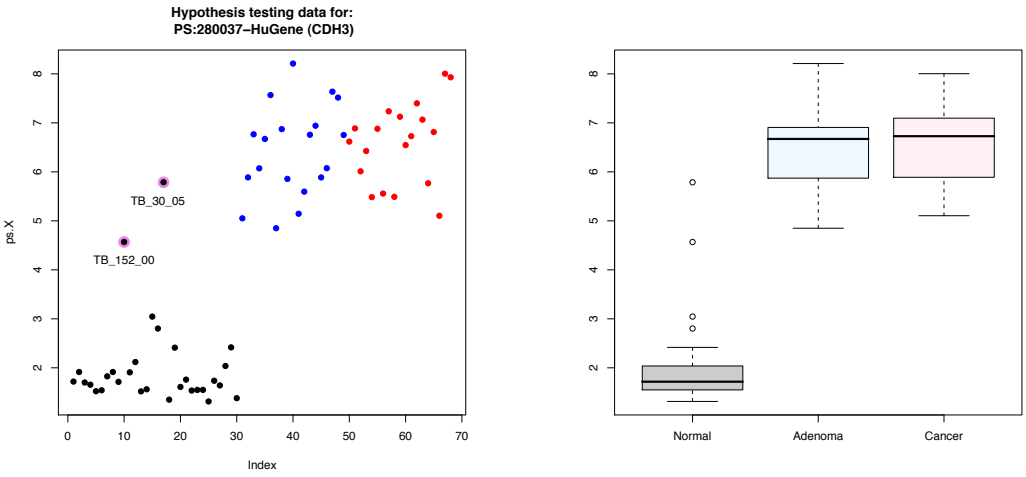


Figure 8.9: Hypothesis testing data for PS:280037-HuGene (*CDH3* The data are shown on the left by tissue and in summary box plot on the right. Note the two outlier "normal" tissues are marked).

8.6.3 Neoplasia specific probesets

In Chapter 7, a hypothesis was proposed that certain probesets exhibiting prototypical "turned-on" and "turned-off" expression patterns could correspond to qualitatively present or absent gene expression transcripts in particular tissues. This hypothesis was based on a novel analysis algorithm involving estimation of a background (off) level probeset intensity threshold. As the custom gene chip did not contain a suitable set of background probesets that could be used to estimate a technical assay background threshold, this method was not applied to these validation data. None the less, visual inspection of many of the confirmed probesets suggests that some, but not all, probesets again exhibited this prototypical on/off expression profile.

Neoplasia-specific probesets from discovery were tested for statistically significant differential expression in the validation data set. Of the 23 probesets which were hypothetically "turned-on" in neoplasia, 20 of these showed significantly increased expression ($P <= 0.05$) in the independent validation data. These validated probesets are shown in Table 8.7. 23 of 35 probesets hypothesised to be "turned-off" in neoplasia likewise showed decreased expression in the validation data. These 23 probesets are listed in the Appended Table D.16, p. 297.

Table 8.7: Probesets which were hypothesised to be exclusively expressed in neoplastic tissues (i.e. "turned-on") which yielded differentially increased expression in the validation data

Probeset ID	Symbol	Fold-Δ(Log2)	t stat	P val (corr)	Likelihood
207850_at	CXCL3	0.73	9.35	1.8715e-12	21.04
209309_at	AZGP1	1.32	8.50	3.1930e-11	17.54
202286_s_at	TACSTD2	1.57	6.93	1.4658e-08	11.08
225806_at	JUB	0.25	6.68	3.0302e-08	10.08
204259_at	MMP7	0.74	5.94	5.0200e-07	7.10
206224_at	CST1	0.98	5.48	2.5357e-06	5.34
241031_at	FAM148A	0.32	5.31	4.2716e-06	4.68
223062_s_at	PSAT1	0.43	5.17	6.3257e-06	4.16
213880_at	LGR5	0.47	4.75	2.5315e-05	2.62
227174_at	WDR72	0.45	4.75	2.5315e-05	2.60
204885_s_at	MSLN	1.28	4.67	3.0665e-05	2.32
219787_s_at	ECT2	0.19	4.25	0.0001	0.86
204475_at	MMP1	0.71	4.18	0.0001	0.64
211506_s_at	IL8	0.99	4.00	0.0002	0.03
222608_s_at	ANLN	0.20	3.95	0.0002	−0.13
214974_x_at	CXCL5	0.15	3.76	0.0005	−0.75
202311_s_at	COL1A1	0.50	2.99	0.0052	−2.97
204702_s_at	NFE2L3	0.20	2.80	0.0085	−3.48
232252_at	DUSP27	0.35	2.72	0.0098	−3.66
226237_at	COL8A1	0.08	2.25	0.0320	−4.76

In another example of this approach, the probeset 235976_at understood to target *SLITRK6*, showed a prototypical "turned-on" pattern in adenoma tissues relative to *both* normals and cancers in the discovery data. Figure 8.10 compares the expression pattern of this *SLITRK6* probeset in the 412 discovery tissues (IBD tissues removed) and also the 68 validation experiments. By inspection, this probeset appears to exhibit a similar elevation in adenomas relative to the other phenotypes both the discovery and validation data.

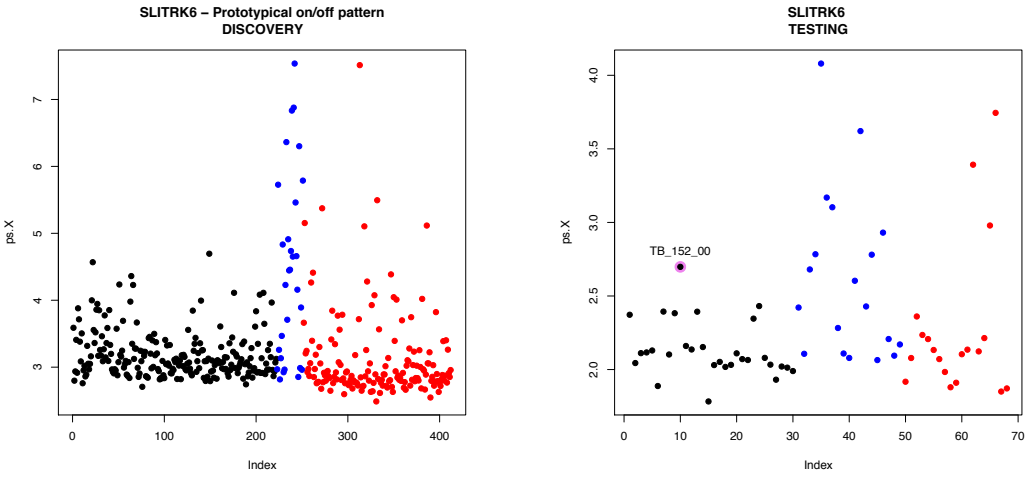


Figure 8.10: Expression profiles for probeset 235976_at in both the Discovery (left) and Validation (right) data sets. Note that the difference between the 19 adenoma and 19 cancer values in the validation data are not significant ($P = 0.0548$). Nonetheless, these expression patterns suggest evidence of an adenoma-specific transcript target. A commonly observed outlier is highlighted in the test data for patient TB_152_00.

8.6.4 Probesets differentially expressed in adenoma versus cancer

Hypotheses related to probesets differentially expressed between adenoma vs. cancer were also tested in the 19 adenoma and 19 cancer tissues. 83 (44%) of the 188 probesets previously discovered to be differentially expressed when compared between cancer and adenoma tissues were likewise differentially expressed in hypothesis testing. These validated probesets for elevation in adenoma or cancer are shown in Appendix Tables D.14 and D.15, respectively. Probesets which target collagen transcripts were conspicuous among the probesets validated to be higher in cancer tissues relative to adenoma, including probesets designed to bind to: *COL4A2*, *COL4A1* and *COL5A2*.

8.7 Hypothesis testing of literature-based candidates

The custom microarray also included 15,114 probesets corresponding to nearly 795 gene symbols that were identified in the literature to be involved in colorectal neoplasia.

6,434 (43%) of these probesets were found to be differentially expressed between the 38 neoplastic tissues (19 adenomas, 19 cancers) relative to the 30 non-neoplastic controls in the test data ($P <= 0.05$) including 4313 probesets expressed higher in neoplasia relative to normal tissues and 2121 probesets lower in neoplastic tissues. Collectively, these probesets are annotated to target 752 gene symbols.

8.8 Candidate biomarkers in common

Finally, differentially expressed genes were compared among all sources of discovery data to assemble a common list of gene symbols (and probesets) which were validated by hypothesis testing. A Venn diagram describing the overlap of confirmed symbols common to the differential display, microarray, and literature lists is shown in Figure 8.11.

8.8.1 Validated genes discovered in this research

There were 22 gene symbols discovered by both differential display research and Affymetrix microarray discovery to be biomarkers up-regulated in neoplastic tissue that were likewise differentially expressed in the validation data. As the differential display discovery data did not include down-regulated markers they are thus not included in the common, or overlapping, lists. The 22 gene symbols which were up-regulated in both the differential display and microarray data are shown in Table 8.8. Furthermore, 14 of the 22 probesets were not observed as

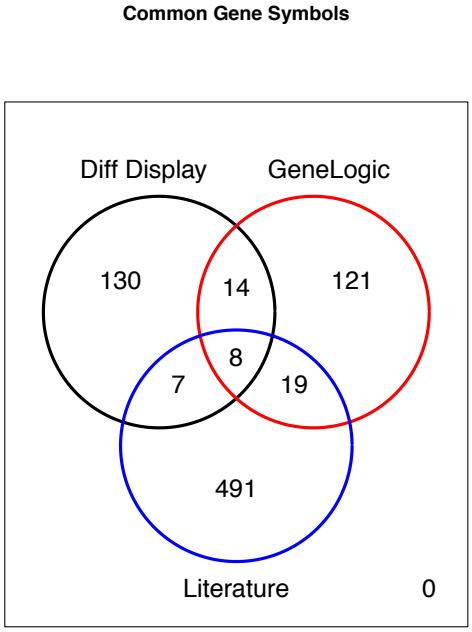


Figure 8.11: Venn diagram describing the overlap of confirmed gene symbols by discovery source.

"literature" markers based on the review carried out *before* these experiments and analyses were conducted by the author and these genes are identified as possibly "novel" colorectal neoplasia markers in Table 8.8. It is important to note, however, that such literature-analyses did not explore e.g. every list of genes demonstrated to be differentially expressed in all published research. Rather, the "literature" list included those genes that were singled out in particular research papers for potential relevance to colorectal neoplasia.

Finally, ROC curves which explore the predictive utility of each of the 14 gene symbols not previously associated with colorectal neoplasia detection (at the time of discovery) were calculated. The top probeset, in terms of predictive utility, was designed to hybridise to S100A11 and is shown in Fig.8.12, and ROC curves for all 14 such "novel" biomarkers are Appended as Fig.4.8,p.298.

Table 8.8: Gene symbols common to both differential display and microarray discovery that have been shown in these hypothesis testing experiments to be likewise differentially expressed between neoplastic and non-neoplastic control tissues.

Novel?	**Probeset**	**Symbol**	***P* value**	**Fold-Δ**	**Sens/Spec**	**CI (95%)**
*	1050447-HuG	S100A11	1.4150e-24	3.83	97.4	93.9-99.1
	897250-HuG	KIAA1199	9.7115e-20	17.44	93.7	88.1-97
*	160440-HuG	SLC12A2	1.1421e-19	2.61	94.3	89-97.4
*	28680-HuG	S100P	9.5129e-19	4.37	93	87-96.6
	195459-HuG	DPEP1	1.1700e-18	23.9	92.6	86.6-96.3
*	680908-HuG	RNF43	1.7574e-18	3.24	92.8	86.9-96.5
*	732854-HuG	GALNT6	4.7442e-17	3.05	91.4	84.9-95.6
	374147-HuG	TGFBI	5.5416e-17	4.31	91.1	84.4-95.3
	22584-HuG	ITGA6	1.2340e-16	2.62	91	84.5-95.3
*	463959-HuG	GPR56	5.6281e-16	2.6	90.2	83.5-94.8
*	323199-HuG	C20orf199	7.9905e-14	2.87	87.2	79.7-92.6
*	445445-HuG	ETS2	3.0373e-13	2.36	86.6	79-92.1
	457141-HuG	IFITM1	1.6954e-12	3.3	85.2	77.4-91
*	238968_at	SLC39A10	2.7115e-12	1.93	85.6	77.8-91.3
*	10322-HuG	PLCB4	1.6362e-10	2.37	82.2	73.8-88.7
*	186424-HuG	REG4	1.8631e-10	7.75	81.7	73.3-88.3
*	216316-HuG	RPESP	2.0245e-10	9.07	81.6	73.1-88.2
*	321418-HuG	NQO1	5.6456e-10	2.25	81.2	72.7-87.9
	546247-HuG	DEFA6	4.7257e-07	7.11	75	65.9-82.6
	1070547-HuG	SPP1	4.1783e-05	3.32	70.3	60.9-78.5
	657765-HuG	REG1A	0.0001	10.93	69.2	59.9-77.5
*	143113-HuG	RETNLB	0.0015	1.47	65.8	56.3-74.5

8.8.2 Biomarkers common to all discovery sources

There were also eight genes that were common to all sources of data, including the literature. These eight genes are shown in Table 8.9.

Figure 8.12: ROC analysis of a prediction model using the S100A11 probeset (HuGene-1050447). This probeset was the best probeset for the best gene in terms of phenotype classification using the validation data.

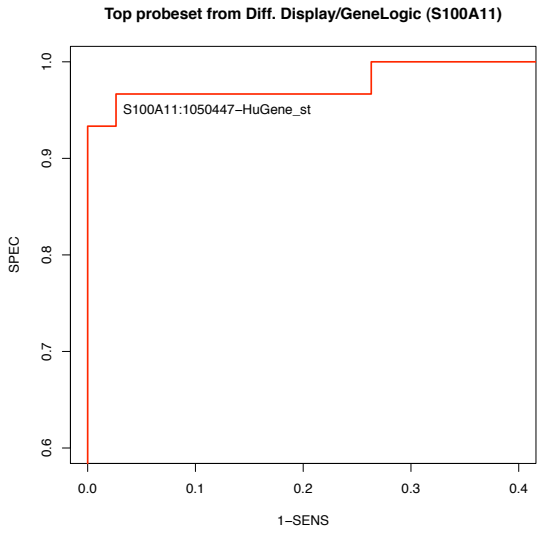


Table 8.9: Confirmed gene symbols that were co-discovered in all data sources.

TGFBI	Transforming growth factor β induced
ITGA6	Integrin, alpha 5
IFITM1	Interferon induced transmembrane protein 1
DPEP1	Dipeptidase 1 (renal)
DEFA6	Defensin alpha 6, paneth cell-specific
REG1A	Regenerating islet-derived 1 alpha
SPP1	Secreted phosphoprotein 1 (osteopontin)
KIAA1199	Unknown, novel gene

8.9 Discussion and conclusions

The primary aim of this research is to identify gene expression biomarkers that will serve as leads for future biomarker research aimed at improving neoplasia diagnosis and screening in the clinical setting. To achieve this aim, two questions are addressed:

1. Are genes differentially expressed in neoplastic tissues relative to non-neoplastic controls?
2. Which of these differentially expressed genes should be chosen as biomarker leads for future study in a clinical context?

The second question necessarily implies that candidate biomarkers must generalise well beyond the current research. This is an important aspect of this project as lack of generalisation often suggests over-fitting hypothetical models to discovery data. This research strives to avoid this problem by applying conservative discovery techniques (e.g. univariate over multivariate approaches) to a relatively large set of data that includes a range of control tissues, including e.g. non-neoplastic disease controls. In addition, new analytical methods are introduced such as the "turned-on" filter which is motivated by the desire to discover fundamental underlying biological transitions associated with neoplastic tissues. The tacit objective of this methodology is that such genes may make more robust biomarkers.

8.9.1 Thesis aim achieved

This chapter describes the validation of candidate mRNA biomarkers by formally testing hypotheses that genes shown to be differentially expressed in discovery analyses are differentially expressed in independently derived clinical specimens. **These results provide evidence that addresses the primary aim of this thesis research by confirming that mRNA transcripts are differentially**

expressed in neoplastic colorectal tissues relative to controls. Approximately 650 genes were validated to be differentially expressed in neoplastic tissues relative to non-neoplastic controls. A complete list of validated genes is Appended in Table D.17, 301. The nature of the discovery data (i.e. the large size, the application of quality control review, the inclusion of non-neoplastic disease controls, etc.) and validation data (i.e. independent clinical specimens) support the conclusions drawn from hypothesis testing.

More importantly, however, these data also address the second question: *"which genes should be chosen for further study?"*. The research introduced a methodology to filter the relatively large list of differentially expressed microarray genes to yield a subset of genes whose expression profile may indicate a qualitative change in expression in neoplastic tissue. Twenty of 23 (87%) of candidates selected for "turned-on" expression patterns were likewise differentially expressed in the validation data as shown in Table 8.7, p. 171. These "turned-on" genes may be particularly useful as candidate leads for future research.

These data do not provide compelling evidence that the use of the "turned-on" filter to identify disease-specific biomarkers produces more robust biomarkers than e.g. stratifying biomarker candidates by other univariate means, such as with Student's t test. The results showed that 387/489 (79%) of adenoma probesets and 440/529 (83%) of the cancer probesets identified in the discovery experiments were validated for differential over- and under-expression (t test, $P <= 0.05$) in the 68 test specimens. While these validation efficiencies (79% and 83%) are slightly lower than the 20/23 (87%) shown for "turned-on" probesets, a closer look at only the adenoma *up-regulated* probesets shows that the validation efficiency between discovery and validation data sets was 103/106 (97%).

In addition to these "turned-on" candidates, another set of 22 genes was observed to be differentially expressed in all sets of data measured in this project. Thus, these 22 genes were: 1) discovered first in a randomly primed differential display experiment in adenoma tissues; 2) identified to be up-regulated in 454 independent clinical specimens measured by 3' biased microarray; and 3) validated in

another set of 68 clinically independent specimens using random primer labeling on oligonucleotide microarray.

A subset of eight genes from these 22 confirm earlier studies carried out by other researchers. The remaining 14 genes are, therefore, relatively "novel" with respect to their potential utility as biomarkers for colorectal neoplasia with respect to the literature based on the review conducted here.

Comparison to the colorectal biomarker discovery literature

As discussed in the review of the colorectal gene expression literature presented in Chapter 2, there is a large and growing literature of gene expression-related experiments carried out on tissues removed from the colorectum [Nannini et al., 2008, Chan et al., 2008]. The data and results of this validation compare well with the literature of colorectal biomarker discovery. In a recent meta-analysis, Chan et al. [2008] describe the concordance of differentially expressed genes across 25 microarray experiments. That review identified five genes to be up-regulated in seven or more independent analyses, including *TGFBI*, *IFITM1*, *MYC*, *SPARC*, *GDF15*. All five of these genes are confirmed to be up-regulated in this study. In particular the top two genes identified in the Chan et al. meta analysis were transforming growth factor-β induced (*TGFBI*) and interferon-induced transmembrane proteins (*IFITM1*). Both of these genes were among the 22 common gene symbols identified in the discovery results reported here. This agreement is all the more interesting because the differential discovery research was aimed at ascertaining the pattern for *adenomas*, not colorectal carcinoma.

TGFBI

TGFBI has been previously shown to be up-regulated in both adenomas and cancers using SAGE technology [Buckhaults et al., 2001, Zhang et al., 1997]. The over-expression of *TGFBI*, which is believed to encode for an extracellular protein involved in cell adhesion [Irigoyen et al., 2008], has been correlated with the increased metastatic potential of colorectal cells [Irigoyen et al., 2008, Ma

et al., 2008]. While apparently up-regulated in colorectal cancer, this gene has been shown to be down-regulated or silenced in human leukemia and cancer cell lines of lung, prostate and colorectum [Li et al., 2009, Shah et al., 2008].

IFTM1

IFITM1 gene expression is induced by interferon gamma and has been shown to increase following Wnt pathway stimulation through β-catenin signalling [Andreu et al., 2006]. Further, over-expression of *IFITM1* has been shown to result in deregulation of cell growth and increased proliferation by stabilizing p53 through phosphorylation inhibition [Yang et al., 2007]. This gene has been previously identified as a candidate biomarker for colorectal neoplasia (including adenomas) and one study has also suggested that anti-IFITM1 antibodies are detectable in serum of 14 of 38 patients with colorectal neoplasia [Liu et al., 2008b, Andreu et al., 2006].

The evidence of increased gene expression in colorectal neoplastic tissues in three separate experiments reported here and multiple previous research publications leads to a well supported conclusion that both *TGFBI* and *IFITM1* genes are potential biomarkers for colorectal neoplasia. In particular, both markers have evidence of increased expression in adenomas as well as cancers. One concern about the clinical utility of these markers, however, is that both biomarkers have also shown evidence of increased expression in cancers outside the colorectum [Hatano et al., 2008, Shah et al., 2008]. See Section 9.6.1 for discussion on the confounding potential of extra-colorectal tumours related to biomarker specificity.

There have been few studies that focus on, or even address, differential expression in adenomas. In this respect this thesis work is a contribution to the field of colorectal neoplasia biomarkers. There are, however, two notable examples, both published recently.

Galamb et al. [2008] measured 20 adenoma specimens, 22 cancer tissues, 11 hyperplastic polyps, 21 IBD specimens and 11 healthy controls using the full genome Affymetrix U133Plus2 oligonucleotide microarray. This research was a

significant expansion of an earlier experiment [Galamb et al., 2006] which also included adenomas but that earlier study used a microarray platform containing a smaller set of genes. In the 2008 study, Galamb et al. applied significance analysis of microarrays to identify a minimum set of three genes differentially expressed in adenomas relative to normal controls (*KIAA1199*, *FOXQ1*, and *CA7*) and a minimum set of five genes to discriminate cancer from normals (*VWF*,*IL8*,*CHI3L1*,*S100A8*, and *GREM1*). A nine gene model was discovered to distinguish adenomas from IBD specimens which likewise included *KIAA1199*. Given the massive expression differences observed between neoplastic and non-neoplastic tissues in the results presented in this thesis, the overlap with the discovery with the results of Galamb et al. is compelling and interesting.

The most comparable study to this work was recently published by Sabates-Bellver et al. who analysed gene expression using Affymetrix U133plus2 microarrays in 32 adenoma specimens and 32 matched normals [Sabates-Bellver et al., 2007]. In that study Sabates-Bellver et al. identified 1,190 up-regulated and 2,469 probesets down-regulated (by mean difference and also 2-fold change) in adenoma tissues relative to matched normals. The list of discovery probesets was not validated except for a small number of selected probesets by RT-PCR.

Sabates-Bellver et al. also identified a subset of 478 (153 up, 325 down) probesets differentially expressed by a four-fold change or more in adenomas. Comparing this list of 153 over-expressed probesets to the 106 probesets discovered in the original 251 microarrays of this research (222 normals, 29 adenomas) yields an overlap of 33 (21.6%) probesets. An additional 20 probesets from the Sabates-Bellver et al. [2007] list were likewise differentially expressed in the cancer vs. normal discovery contrast bringing the total number of overlapping probesets to 53, or 34.6%. This is a high level of correspondence.

KIAA1199

The Sabates-Bellver et al. research was particularly focused on the differential expression of Wnt-related genes. The results reported here are in agreement with their conclusion that Wnt-related genes are differentially expressed in a high proportion. In addition, Sabates-Bellver singled out the gene *KIAA1199*

as a novel target of the Wnt pathway and a possible novel biomarker for colorectal neoplasia. *KIAA1199* was likewise highly differentially expressed in both discovery data sets and the validation data of this thesis. At the time of this report *KIAA1199* remains a novel gene of unknown function or structure, although Sabates-Bellver showed that this gene appears strongly correlated with other Wnt-related genes to a significant degree [Sabates-Bellver et al., 2007]. Nevertheless, I conclude from these data that *KIAA1199* is differentially expressed in both adenomatous and cancerous polyps and is a worthy target for future research.

Neoplasia biomarker panel

Twenty-two genes were discovered in common between the two discovery data sets and also validated in a third data set (See Table 8.8). Ten of these 22 validated biomarkers demonstrated a high sensitivity and specificity in the validation data set of over 90%, including both *KIAA1199* and *TGFBI* (*IFITM1* was 85%). These remaining eight biomarkers, which demonstrated a sensitivity and specificity above 90%, are discussed here.

S100A11

The strongest biomarker candidate identified in this research in terms of neoplasia vs. non-neoplastic discrimination *in the validation data* was *S100A11*, also known as calgizzarin or S100C [Reichling et al., 2005]. In the validation data, *S100A11* mRNA transcripts (Sequence ID 302), demonstrated a 97% sensitivity and specificity (See Figure 8.5). *S100A11* is a member of the S100 super-family of Calcium binding EF-hand motif proteins which includes 20 members. As a family, these proteins are known to be involved with a wide range of cell functions and *S100A11*, in particular, has been shown to regulate cell proliferation [Salama et al., 2008].

While *S100A11* was included in the list of adenoma biomarkers identified by Sabates-Bellver [Sabates-Bellver et al., 2007] and earlier studies of APC^{min} tumours [Tanaka et al., 1995], this gene has also previously been shown to exhibit

a range of activity across other cancers. *S100A11* has been reported to exhibit tumor promoter activity in breast and prostate cancers but tumour suppressor activity in renal and bladder cancer [Salama et al., 2008]. This gene has also been shown to be over-expressed in breast, lung squamous cell cancer, lung adenocarcinoma and renal cell cancer by subtractive hybridisation and microarrays [Amatschek et al., 2004], but down-regulated in leukemia [Li et al., 2009]. Interestingly, *S100A11* has also been demonstrated to be down-regulated in response to administration of mitomycin C and 5-fluorouracil in biopsy specimens from patients with rectal cancer.

More recently, protein-level experiments using SELDI-based mass-spectroscopy have shown that *S100A11* can be used to cluster and distinguish metastatic tumours originating from colorectal and hepatocellular primary tumours [Melle et al., 2008].

Based on the currently available literature, there appears to be ambiguity about the precise function of *S100A11* in the colorectal mucosa and also it's involvement in a broad spectrum of cancers. Nevertheless, the gene expression data presented here confirm earlier findings [Sabates-Bellver et al., 2007, Tanaka et al., 1995] and convincingly support the conclusion that this gene is over-expressed in colorectal neoplasia compared to non-neoplastic controls.

SLCA2

SLC12A2 was also up-regulated in this and previous studies [Habermann et al., 2007, Sabates-Bellver et al., 2007, Notterman et al., 2001, Takemasa et al., 2001, Bertucci et al., 2004, Ohmachi et al., 2006, Seiden-Long et al., 2006]. This gene is one of nine members of the SLC12 family of cation coupled chloride co-transporters. The major function of the protein encoded by this gene in epithelial cells is to provide the cell with Cl^{-1}, which is then secreted. Disruption of this gene has been observed in several human diseases including inner-ear dysfunction, a defect in spermatocyte production, reduction in saliva, and sensory perception abnormalities [Hebert et al., 2004].

SLC12A2 is believed to be a downstream target of Wnt signalling [van de We-

tering et al., 2002] and the gene has been shown to be induced in colorectal cell lines by stimulated hepatocyte growth factor [Seiden-Long et al., 2006].

Of particular interest is the observation by Habermann et al. [2007] that while *SLC12A2* is over-expressed in both adenomas and cancer tissues relative to normal tissues, the expression of this gene *drops* in cancer tissues relative to adenomas. Five of 24 probesets in the validation which are designed to hybridise to *SLC12A2* mRNA clearly agree with the findings of Habermann et al., with very significantly lower expression levels in the 19 cancer tissues relative to the 19 adenoma tissues for both the conventional (3' biased) U133plus2 probesets and exon-based HuGene probes. Given this gene's particular over-expression early in the adenoma-carcinoma sequence, *SLC12A2* should be included in future biomarker studies.

S100P

S100P is the second member of the S100 family to be shown in this research to be a useful biomarker candidate for colorectal neoplasia [Salama et al., 2008]. This gene was likewise identified by Sabates-Bellver et al., who observed this gene to be more than four-fold increased in adenoma tissues, in agreement with other colorectal cancer gene expression analyses [Sabates-Bellver et al., 2007, Bicciato et al., 2003, Datta and Datta, 2005].

S100P protein is believed to stimulate intracellular signalling cascades after binding to "receptor for advanced glycation end products" (RAGE), a protein which may mediate colitis through activation of NF-κB signalling [Turovskaya et al., 2008].

In addition to over-expression in neoplastic tissues, however, the results of this thesis suggest that *S100P* is expressed lower in *normal* proximal tissues relative to normal distal tissue. The validation data used in this research were well balanced with respect to specimen location and the neoplastic vs. non-neoplastic comparison are convincing in both data sets. Nevertheless, the evidence of proximal vs. distal expression changes in the non-neoplastic phenotype argues caution in the use of *S100P* as a biomarker candidate. *S100P* does not appear

to be a preferred biomarker based on these collective observations.

DPEP1

Renal di-peptide peptidase 1 (*DPEP1*) is a membrane-bound glycoprotein involved in di-peptide hydrolysis in the kidney [Nitanai et al., 2002]. *DPEP1* was the most differentially expressed gene identified by Ohmachi et al. [2006] using a 12,800 gene cDNA microarray analysing 16 colorectal cancer patients, confirming earlier SAGE evidence of *DPEP1* over-expression in colorectal neoplasia [Huang et al., 2006]. Using a new sequence tag-based technique called massively parallel signature sequencing (MPSS), *DPEP1* was observed to be expressed in colorectal cancer tissues compared to weak expression in normal tissue [Alves et al., 2008]. This evidence prompted Alves et al. to hypothesise that *DPEP1* could be neoplasia specific in an analogous manner to the concept of neoplasia specific expression suggested here [Alves et al., 2008]. Nevertheless, *DPEP1* did not fit the "turned-on" pattern in these data.

DPEP1 is also interesting as several studies have demonstrated neoplastic differential expression in human blood and faecal specimens. In addition to showing over expression of *DPEP1* in colorectal tissues, McIver et al. [2004] also detected *DPEP1* mRNA in the peripheral blood of 15/38 cancer patients by RT-PCR. Finally, DPEP1 was also discovered to be among three genes differentially expressed in cancer patients by assaying colonocytes isolated from human stool [Yajima et al., 2007]. Interestingly Yajima et al. demonstrated that high quality RNA and DNA can be isolated from human stool using a combination of filtration and magnetic-based cell sorting. By using Affymetrix GeneChip discovery and RT-PCR validation, Yajima et al. concluded that *DPEP1* is a useful candidate for detecting cancer of any stage.

Based on the data collected here and the intriguing suggestion that this gene is detectable in blood and stool of colorectal cancer patients, *DPEP1* is suggested as a biomarker candidate for colorectal neoplasia.

RNF43

Ring finger protein 43 (*RNF43*) was first described in 2004 by Yagyu et al.

[2004] as a novel human gene over-expressed in colorectal cancer tissues [Yagyu et al., 2004]. While the protein remains relatively poorly characterised, *RNF43* is believed to exhibit ubiquitin ligase activity due to the presence of the RING finger domain, although this activity has not yet been confirmed biologically [Sugiura et al., 2008]. On the other hand, inducing gene expression of *RNF43* has been shown to exert growth promoting activity while gene knockdown by RNA interference resulted in growth suppression [Yagyu et al., 2004].

The results of the discovery and validation experiments described here suggest that *RNF43* should be included in this panel of candidate biomarkers.

GALNT6

GALNT6 is a member of the UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase enzyme family which are generally involved in mucin type O-linked glycosylation. This member (along with *GALNT3*) is capable of glycosylating fibronectin to form the foetal antigen, glycosylated onco-foetal fibronectin, which is deposited in tumour stroma in some cancers such as oral squamous cell carcinoma [Bennett et al., 1999, Wandall et al., 2007]. The gene has been previously shown to be differentially expressed in colorectal adenomas [Sabates-Bellver et al., 2007]. The gene has also been shown to be over-expressed in breast cancer at both the gene and protein levels [Berois et al., 2006, Freire et al., 2006].

GALNT6 is an interesting biomarker candidate because of its presumed role of generating foetal glycosylation patterns. Further the consistent discovery of this gene in the adenoma-focused differential display discovery, the microarray discovery data, and also the adenoma vs. normal contrast carried out by Sabates-Bellver et al. suggest that this gene is a reasonable biomarker for consideration with sensitivity for adenomas. On the other hand, as with *TGFBI*, *GALNT6* has also demonstrated extra-colonic cancer activity [Berois et al., 2006, Freire et al., 2006].

Given the possible role of *GALNT6* in constructing an onco-foetal stromal architecture, this gene is advanced by this research for future clinical validation. Nev-

ertheless, the possibility of non-neoplastic causes of over-expression also highlights the need for careful specificity studies.

ITGA6

ITGA6 is a member of the of the alpha subunit integrin family which is generally involved in cell-cell adhesion and signalling by binding extracellular glycoproteins (e.g. laminins) [Georges-Labouesse et al., 1998, Segditsas et al., 2008]. Like *SLC12A2*, *ITGA6* has been shown to differentially expressed in response to hepatocyte growth factor [Seiden-Long et al., 2006]. This protein may play an essential role in maintaining proliferative and growth potential in tumour cells. Cariati demonstrated that *ITGA6* was essential for growth and survival of the breast cancer cell line MCF-7 and that gene knockdown by RNA interference results in reduced tumourigenicity in mice [Cariati et al., 2008].

This gene has previously been shown to be up-regulated in colorectal adenocarcinoma and also oesophageal cancer by cDNA microarray [Chen et al., 2006, Hourihan et al., 2003]. Using *in situ* hybridisation *ITGA6* has been shown to exhibit a uniform, diffuse pattern in colorectal adenomas, while no expression was observed in normal control tissues [Segditsas et al., 2008]. Finally, in a small study of just five tumours with matched normals, Kim et al. [2008b] observed that *ITGA6* was one of the top five most differentially expressed genes (by fold change) in a serrated adenomas.

The suggested over-expression of *ITGA6* in serrated adenomas suggests possible involvement in a non-Wnt pathway. Segditsas et al. [2008] likewise suggested that this gene may not be a direct target of Wnt activation based on lower levels of expression change compared to other Wnt targets.

Based on the *ITGA6* expression results of this work and the potential that the gene could provide possible diagnostic utility for broader spectrum of colorectal carcinoma (i.e. non-Wnt perturbed), this gene is included in the panel of biomarker candidates.

GPR56

GPR56 (previously identified as TM7XN1) is a member of the largest family

of cell-surface receptors, GTP binding protein receptors, and *GPR56* is over-expressed in a range of human cancers [Liu et al., 1999, Huang et al., 2008]. *GPR56* has been measured to be differentially over-expressed in oesophageal cancer, including early tumours, by RT-PCR compared to matched normal [Sud et al., 2006] and also glioblastoma multiforme tumours [Shashidhar et al., 2005].

Recently, Xu and Hynes suggested that *GPR56* is a suppressor of tumour progression and metastasis through its interaction with the transglutaminase TG2 [Xu and Hynes, 2007]. TG2 is believed to cross-link cell surface proteins (such as fibronectin, integrins, and GPR56) to extracellular matrix proteins, which may prevent access to matrix proteinases such as those of the MMP family, thus inhibiting tumour migration.

Interestingly, this thesis appears to be the first study to identify differential expression of *GPR56* in colorectal neoplasia. Given this novel observation and the intriguing role of *GPR56* possible relation to other biomarkers suggested here (e.g. the integrin *ITGA6*) this gene is included in the panel of candidate biomarkers for future study.

8.9.2 Conclusion

The aim of this work is to identify candidate biomarker leads for future assay development and research, not to uncover a biological rationale for gene expression differences in neoplasia. Nevertheless, as exemplified by the review of these ten genes, a biological understanding of biomarkers provides potentially important diagnostic application context. For instance, some genes are more likely than others to be expressed in tissues outside the colon in either the disease or healthy state. Finally, understanding the potential for markers to represent divergent pathways for carcinogenesis (e.g. serrated polyp pathway vs. adenoma-carcinoma sequence) likewise broadens the sensitivity for a panel of candidate biomarkers for heterogeneous disease such as colorectal neoplasia.

Chapter 9

Conclusions

9.1 Overview

At the time of this writing, there is significant debate about the validity of using gene expression microarray data to predict patient phenotype. The key points of this debate relate to the general lack of high-quality validation or test experiments and the dangers of overfitting predictive models to training data [Ioannidis, 2005]. This thesis describes the discovery of a set of RNA transcript targets that have been experimentally tested with a high degree of rigor. The number of clinical specimens measured for both discovery and hypothesis-testing are each relatively large compared to many studies in the literature.

The central aim of this thesis is to identify candidate biomarkers of colorectal neoplasia for future assay development and testing. These data provide convincing support for the following two conclusions in respect of this aim.

1. I conclude that there are genes which are differentially expressed between neoplastic colorectal tissues and non-neoplastic controls in a consistent, robust manner. This conclusion agrees with the literature where previous studies have also demonstrated differentially expressed genes between colorectal phenotypes [Chan et al., 2008]. However, while this research finding is not novel, the study presented here utilises larger discovery and validation experiments than pre-

viously published. Further, the application of multivariate techniques such as PCA to the full human transcriptome as measured by oligonucleotide microarrays provides evidence of the overwhelming affect of the neoplastic phenotype on gene expression variability. These data are perhaps the first to provide evidence of simultaneous phenotype discrimination of normal colorectal tissue, IBD specimens, adenomas and cancer using only a gene expression data set.

2. Having convincingly established that some genes exhibit gene expression patterns which correlate with the neoplastic phenotype, these results enable investigation of the central practical aim: the identification from the pool of differentially expressed genes those candidate biomarkers which could serve as leads for clinical assay research and development in the future. For example, the validation results reported herein demonstrate that single marker logistic regression models constructed using just one probeset can achieve up to 97% correct classification of a relatively large number of specimens (38 neoplastic tissues, 30 non-neoplastic controls) (See S100A11 probeset values, p.175). Multiplexing small subsets of probesets achieves perfect discrimination of these phenotypes. These univariate and small-panel multivariate results are generally stronger than what has been previously reported in the literature in terms of biomarker validation. This work thus establishes sufficient *in vitro* evidence to warrant progressing, as proposed by Pepe et al., the best candidates to future research with the new aim of developing *in vitro* assays to diagnose colorectal neoplasia [Pepe et al., 2001].

9.2 Analysis of gene expression microarrays

Univariate vs. multivariate results

The promise of gene expression-based biomarker discovery and the expectation to apply these technologies to clinical diagnosis is well documented [Ransohoff, 2004b]. Complex gene expression-based diagnostic and prognostic studies have been suggested for many forms of cancer [Alizadeh et al., 2001, Tinker et al.,

2006]. Unfortunately, the promise of these suggestions has yet to be realized; no new biomarkers have recently been approved for colorectal neoplasia diagnosis, and there is a growing body of literature highlighting the problems of complex diagnostic models based on e.g. gene expression, proteomic fingerprinting, etc. [Nannini et al., 2008, Soreide et al., 2008, Sotiriou and Piccart, 2007, Shi et al., 2006, Ransohoff, 2004b].

The most serious – and perhaps the most common – difficulty related to biomarker discovery is overfitting a complex model to a limited discovery data set which leads to a lack of generalisability of the resulting model [Tinker et al., 2006, Ransohoff, 2004b, Hastie et al., 2001]. The risks of overfitting can be mitigated by a range of analysis techniques, some of which are discussed in this work, including e.g. penalized learning methods that aim to reduce model complexity [Hand, 1997] and subset selection (See Chapter 4) which aims to limit the complexity by lowering the number of model parameters. Surprisingly, the data reported herein demonstrated that some single probesets (or transcripts in the case of RT-PCR) provide relatively strong discriminating power between colorectal neoplastic specimens and non-neoplastic controls. Many of these biomarkers show similarly strong discrimination power in classifying these phenotypes in the independent validation data.

The opportunity to utilise single biomarkers to predict colorectal neoplasia could greatly simplify the future work required to formulate a diagnostic *in vitro* assay for clinical use [Pepe et al., 2001]. Nevertheless, the results of this thesis have also shown that combining several marginally effective univariate biomarkers using, for example, logistic regression models greatly improves tissue classification results.

The discovery of numerous examples of univariate biomarkers with strong classification efficiency in both the discovery and validation experiments enabled this research to follow a simpler, perhaps more convenient, research direction. There is a growing field of statistical learning literature aimed at addressing the mathematical problems of analysing high dimensional data sets such as microarray data. The central difficulty of these analyses is choosing among models when

the number of features greatly exceeds the number of observations. While many sophisticated techniques have been suggested, any method of discriminating phenotypes involving multiple variables necessarily requires a larger validation data set compared to univariate solutions. Unfortunately, the number of required validation tissues quickly increases with the dimension of the model. On the other hand, simple univariate solutions afford a more confident appreciation of variance within and between phenotypes.

Furthermore, a relatively limited set 71 observations measuring 67 RT-PCR targets showed that sophisticated multivariate techniques often 're-discover' the strongest classifiers from univariate analysis. These limited data also suggest, however, that relatively poor univariate features are sometimes recruited into discovery models to improve multivariate classifiers for specific observations that were otherwise misclassified by the simpler models.

Nevertheless, given the heterogeneous nature of most diseases, including colorectal neoplasia, I anticipate that improved diagnostic utility will likely be achieved by combining these univariate markers into multivariate models. There is evidence of this improvement in the data presented here, but this view will require further testing before a conclusion is warranted. In the mean time, this thesis offers a number of compelling univariate solutions to discriminate colorectal neoplasia from non-neoplastic controls.

Identification of phenotype-specific RNA transcripts

Most biomarker discovery research, including the main body of this work, attempts to discover differentially expressed features (in this case, RNA transcripts) based on a quantitative change between phenotypes of interest. Commonly used metrics for establishing quantitative difference levels include fold change [Yang et al., 2002] and differences in means using t-tests [Comander et al., 2004, Smyth, 2005]. This thesis introduces an alternative analysis technique that is motivated by an aim to shift the diagnostic interpretation of a putative biomarker from precise quantification to a simpler 'present' or 'absent'

decision (See 5.5.9, p.94). This method seeks to direct biomarker discovery to a subset of biomarker targets that exhibit a qualitative change in expression between phenotypes.

Applied to the microarray discovery data, this methodology identified a subset of the probesets which appear to exhibit such a qualitative expression change associated with neoplasia (See 7.3.4, p.135). These particular patterns suggest the possibility that these transcripts are transcriptionally silenced ("turned-off") in one phenotype, but expressed at a detectable level ("turned-on") in a second phenotype. Putative biomarkers that are transcriptionally absent in non-neoplastic tissues, but positively expressed in neoplasia, could reflect transcription events specific to colorectal adenoma (and carcinoma) formation. Further, these neoplasia specific transcripts could potentially be translated into proteins which are likewise neoplasia specific. Experience developing commercial *in vitro* diagnostic assays suggests that reporting protein analytes as being "present" or "absent" could greatly simplify the assay development process by avoiding quantification, with its attendant requirements for standards, etc.. These patterns may provide an opportunity to create high sensitivity assays which discriminate neoplastic from non-neoplastic specimens based on the simple presence or absence of one or more of these biomarkers.

While the implementation introduced in this thesis is perhaps simplistic, the methodology could be refined, for example, by introducing more sophisticated modeling to predict those markers which are transcriptionally-absent in a phenotype-specific manner. In particular, visual inspection of those biomarkers which appear to exhibit a qualitative expression profile between phenotypes suggests that the variability in "off" tissues is lower compared to the "on" tissues. One explanation for this observation could be that, in the "off" state, the primary source of variance among gene expression measurements is technical variability, while the "on" genes exhibit both technical *and biological* variability.

The validation data utilised in these experiments was not suitable for testing hypotheses in respect of biomarkers which are "turned-on" or "turned-off" because of the nature of selected genes included on the custom microarray. On

going experiments which are beyond the scope of this thesis, however, suggest that some mRNA transcripts appear to be neoplasia specific.

The utility of gene set enrichment analysis

Gene set enrichment analysis (GSEA) is an analysis algorithm developed to improve the reliability of microarray discovery [Subramanian et al., 2005]. By identifying gene-expression differences across *a priori* defined gene pathways, this method aims to distill broad underlying gene expression changes without focusing on data at the level of individual transcripts [Efron and Tibshirani, 2006]. The methodology uses a system-biology level approach which examines expression at a pathway-level instead of the more conventional gene-level expression analysis.

GSEA was motivated by the aim to lessen the impact of inter-experiment variation that can lead to poor reproducibility and unsuccessful validation [Subramanian et al., 2005]. In addition, there is the possibility that even a small perturbation in a group of genes may be detectable by gene set enrichment even if individual gene changes within the set are non-significant. Despite these usual concerns, however, the correspondence and reproducibility of gene-level findings in the discovery and validation data reported here work was high, even using independently collected tissues and varying expression measurement technologies. Nevertheless, the results also show that GSEA can usefully be applied to identify correlates between genome-wide gene expression and neoplastic phenotypes. GSEA comparing normal colorectal tissues and inflamed tissues (colitis) highlights the altered binding of immune-response related probesets between these phenotypes. GSEA was also employed to demonstrate that the Wnt-target genes are significantly perturbed in neoplastic tissues relative to non-neoplasia controls. Finally, a comparison of the Wnt target genes taken from the publicly available KEGG database versus a hand-curated list suggests the importance of careful gene list construction. While the publicly available KEGG list of Wnt-related probesets was only marginally informative, a list of manually assembled

TCF/Lef gene targets determined from the literature was significantly altered in neoplasia (See Section 7.4.1, p.144).

In conclusion, GSEA gave expected results for inflammation-associated genes in IBD and Wnt signalling pathways in neoplasia. These data provide confidence that this methodology may be capable of providing biologically relevant results.

The utility of PCA to visualize high dimensional data

Results reported herein demonstrate that multivariate analyses play a role in all aspects of the biomarker discovery process, including quality control, predictor discovery and hypothesis testing. The ability to examine and compare a set of observations across a wide number of features is particularly useful to observe high level trends in the data. PCA (unsupervised and supervised) was applied here to create two-dimensional projections of high-dimensional data sets which highlight relationships between observations and phenotypes.

PCA was usefully applied in quality control analyses to identify a subset of tissues that were processed differently (e.g. micro-sample amplification) than the bulk of the data. While analysing the discovery microarray data, PCA raised concerns about potentially confounding variables involving a subset of tissues that showed histological evidence of a substantial muscularis contamination in the discovery data. A careful subsequent review of the histology description of those tissues confirmed a set of observations contaminated by muscularis tissue. After isolating and scrubbing these observations from the data, the underlying genome-wide neoplasia vs. normal phenotype relationships between the observations came into sharp focus.

A PCA plot of these validation data illustrates that probeset selections made for the custom chip design are useful for three-class discrimination of normal, adenoma, and cancer tissues (See 8.4, p.160). Further, a PCA of the microarray discovery subset using only those probesets believed to be involved in the Wnt signalling pathway provides the first evidence of four-class discrimination between normal, colitis, adenoma and cancerous tissues.

Finally, a version of principal component analysis designed to be robust to the influence of outliers was recently introduced [Hubert et al., 2005]. Given the relatively large role PCA has played in this thesis, this robust method should be explored. In preliminary application to the entire discovery data the robust PCA algorithm produced the characteristic "neoplasia vs. non.neoplasia" clusters as shown in Figure 7.6, 136. Surprisingly, the robust method was also able to distinguish these two phenotype clusters even before a set of 28 tissues contaminated with muscularis were removed (data not shown). This result highlights both the strength and weakness of using the robust method. The contaminating tissues were not likely to be identified using the robust PCA method alone, suggesting a limitation for quality control purposes and potentially confounding the results of this research. On the other hand, one should also recognize that it is not possible to predict, or test for, every possible confounding variable. Recognizing this fact, the robust PCA may provide a method to safeguard ourselves against potential bias from unrecognized outliers while still providing a representative principal component plot.

In conclusion, multivariate visualization techniques such as principal components analysis provide valuable insights about high level trends between observations.

Critical impact of quality control

A set of novel analytical tools and a methodology for assessing the overall internal consistency and conformity of microarray data was developed and applied here [LaPointe and Dunne, 2005a]. These methods augment published quality control metrics with new techniques such as slope analysis of degradation plots for 3' biased arrays, inter-chip comparisons for U133A/B data and principal components analysis to identify confounding variables. As discussed above, the ability to recognize and remove chips processed using tissue samples contaminated with muscularis significantly improved neoplasia vs. non-neoplasia resolution in the discovery data.

I conclude that the quality control techniques introduced in this work are useful for understanding potentially confounding effects which lead to outlying observations.

9.3 Gene expression along the normal colon

Value of understanding normal gene expression patterns

Overfitting a prediction model to discovery data may result in poor downstream validation performance. An equally serious and related difficulty is the failure to protect against confounding experimental bias. A review of the biomarker discovery literature suggests that, for most experiments, there is little, if any, attention paid to understanding the full range of normal variability. Further, most gene expression discovery reports fail to address the potential impact of non-disease related gene expression patterns on the data [Pepe et al., 2001].

Chapter 6 identified and characterised gene expression patterns which occur in non-diseased colorectal tissue along the longitudinal axis of the organ. These expression patterns include individual transcripts which undergo highly significant, multiple-fold, increases or decreases between the proximal and distal large intestine (See Section 6.3.2, p.106). These (published) results support the conclusion that failure to understand the potential for anatomy-specific expression patterns of such transcripts could significantly confound the biomarker discovery process in diseased tissues.

More generally, the obvious, but critical, observation is made that discovery of reliable, robust disease-specific biomarkers must be based on a thorough understanding of the range of expression patterns in control tissues. Where possible, the analysis of control tissues should include unrelated disease specimens of the same organ [Pepe et al., 2001]. In microarray discovery data used for this work, there were 42 non-neoplastic colitis control specimens. Results of gene set analysis using GSEA confirm that these tissues exhibit significantly

altered immune-response gene expression patterns (See Section D.4.5, p.287). Such immune-response pathways could potentially be involved in host-tumour response and might therefore be "discovered" by discriminant techniques. In fact, markers of host-response to colorectal neoplasia *may* be useful as diagnostic markers. Such markers, however, would obviously be non-specific for neoplasia in a clinical context, and they should be carefully considered before inclusion to a candidate diagnostic panel.

Influence of colorectal location on gene expression

The large intestine is typically segmented into six anatomical regions: caecum, ascending colon, transverse colon, descending colon, sigmoid colon, and rectum [Yamada et al., 2003]. Several studies report differential distribution of diseases and their incidence between the proximal and distal colon [Bufill, 1990, Distler and Holt, 1997]. This evidence for functional and pathological differences between the proximal and distal colorectum suggests the question of whether the underlying gene expression patterns vary between the different regions?

Using genome-wide microarrays, 115 probesets (out of 44,928) are differentially expressed between the terminal segments of the caecum and the rectum and 206 probesets (corresponding to approximately 150 genes) are differentially expressed between the proximal and distal segments in aggregate (See Section 6.3.2, p.106). These observations suggest that from a genome perspective only a small number of genes are differentially expressed. For example, 206 probesets represents only 0.5% of the 44,928 probesets tested in these experiments.

I therefore conclude that gene expression varies along the colon but the relative impact, in terms of the number of probesets that change from the proximal to distal colon, is not large.

How do genes change longitudinally?

There are two gene expression profiles evident in the 115 probesets differentially expressed between the caecum and rectum. The first pattern is consistent with a two tier proximal vs. distal model of expression change. The second pattern suggests a multi-segment model of more gradual change moving distally (See 6.3.3, p.110). In the discovery data, the first pattern is represented by 65/115 probesets, while the second pattern is observed in the remaining 50/115 probesets. A supervised principal components analysis in the subspace of only the 115 differentially expressed probesets also yields two clusters of the tissues corresponding to a first cluster of proximal tissues and a second cluster of distal tissues.

I conclude that the dominant pattern of gene expression change along the colorectum correlates with a two-tier proximal vs. distal view of the data. Further, expression of these differentially expressed genes often changes abruptly at the transition between the ascending and descending colon. A smaller number of probesets exhibit a gradual inter-segment expression change moving distally between segments.

Intrinsic vs. extrinsic expression patterns

How might one explain these two models? Examination of the differentially expressed probesets along the large intestine reveals an intriguing trend. Of the probesets that exhibit a sharp increase change between the proximal and distal transitions (i.e the majority of differentially expressed probesets), approximately half are elevated in the proximal tissues and half are elevated in the distal tissues. For the *gradually* changing probesets, approximately 90% of the probesets show increasing expression between the segments proceeding distally from the proximal to distal segments.

This dominant proximal vs. distal expression pattern correlates well with the predicted embryological midgut vs. hindgut patterns established during em-

bryogenesis [Babyatsky and Podolsky, 2003]. I hypothesise that this dominant expression pattern therefore reflects the intrinsic underlying ontogeny of the large intestine. The sharp model change between these tissues and also the balanced numbers of increasing and decreasing probesets support this hypothesis.

There is also a secondary pattern evident that exhibits a gradual changing expression pattern increasing from proximal to distal segments. As increase of these genes is in the same direction as the flow of luminal contents through the gut, this pattern might be explained by 'environmental' changes induced by differential patterns of luminal content and events along the length of the colon. These environmental changes could include differential flow of food stuffs from the small intestine and progressively changing microflora and metabolism of luminal substrates such as carbohydrate and protein fermentation [Macfarlane et al., 1992]. The later fermentative events are well known to a show a differential longitudinal pattern that is variably affected by diet.

In conclusion, there is significant variation of gene expression in approximately 200 genes along the colorectum. Two distinct patterns of variation are observed among these genes. One pattern is a bidirectional proximal-distal change that is abrupt and fits with ontological development. The secondary pattern is one of gradual change where most (90%) of genes involved increase moving distally. This pattern might be explained by environmental regulation.

9.4 Neoplastic gene expression in the colorectum

Design and validation of the custom microarray

A custom gene chip was designed and fabricated to test the hypotheses generated during biomarker discovery (See 8.2, p.155). This custom microarray provides several advantages. First, the custom microarray is a useful tool to simultaneously measure the full set of RNA transcripts that were discovered using differential display of the random-primed transcriptome. Many of those

candidates are not represented by oligonucleotide probesets on commercially available microarrays. Also, by using a randomly-primed labeling technique for RNA extracts, these experiments are not restricted to a 3' biased transcriptome as with traditional Affymetrix microarrays. The random-hexamer based procedure provides access to individual probes across the full open reading frame (ORF) of each target gene. The ability to conveniently measure multiple targets within the ORF of candidate targets, such as by exon-level analysis provides important biological information. In continuing work that is beyond the scope of this thesis, we are beginning to appreciate that a complete understanding of the exon-level expression for each candidate biomarker is important to identifying precisely targeted, disease specific biomarker candidates. For example, probe level analysis against several of the best candidates in these data suggest that there may be evidence of alternative splice processing in neoplastic tissue (data not shown).

Transcript expression trends

In both microarray experiments described here the number of probesets (and putative genes) exhibiting lower expression in neoplasia relative to controls is approximately three times higher than the number of probesets elevated in neoplasia. This observation is consistent with the literature (See Table A.2, p.227).

Neoplasia phenotype and gene expression

The presence (or absence) of the neoplasia phenotype correlates with the largest source of genome-wide variance observed in the discovery data of 454 microarrays (190 neoplastic specimens, 264 non-neoplasia controls). Approximately 25% of the probesets on full-genome microarrays are differentially expressed between neoplastic tissue and non-neoplastic controls, even using highly conservative estimates of mean difference (See Section 7.3.3, p.132). All other phenotype contrasts (e.g. colitis vs. normal, adenoma vs. cancer) resulted in many fewer probesets which were differentially expressed.

Wnt expression pattern

The Wnt expression pathway is reported to be perturbed in over 90% of colorectal neoplastic tissues [Klaus and Birchmeier, 2008, van Leeuwen et al., 2006]. Consequently, one would expect to observe significant expression changes in genes whose transcription is modulated by the Wnt pathway. Indeed, the data provide two elements of strong evidence for Wnt-related effects in colorectal disease. First, there is a significant group-wise expression increase in neoplastic tissues relative to non-neoplastic controls of probesets which bind to putative gene targets of the Wnt pathway. In particular, this group of probesets was the most differentially expressed pathway observed (based on KEGG-derived gene lists) between adenomas and non-neoplastic controls. This observation is consistent with the literature that aberrant Wnt is involved with adenoma formation.

Additionally, supervised PCA plots using only the Wnt-target probesets provides the most compelling phenotype-specific clustering (using four phenotype classes: cancer, adenoma, IBD and normal) of all gene lists tested during this research (See Section 7.4.1, p.144).

9.5 Biomarkers for colorectal neoplasia

Two rounds of biomarker discovery using first differential display of the adenoma transcriptome and then genome-wide oligonucleotide microarray provided this project with a high number of candidate biomarkers for colorectal neoplasia. Validation experiments aimed at testing these candidates in an independent set of clinical specimens confirmed that many of these biomarkers are indeed differentially expressed in neoplastic tissues relative to non-neoplastic controls.

Such a massive expression difference between these phenotypes presents a large number of biomarker candidates for evaluation. The challenge is not, however, to identify biomarkers that discriminate phenotypes in these 454 tissues – which are plentiful in these data – but rather to discover the most robust biomarkers that

will survive downstream hypothesis testing, product development, and clinical validation and so be useful for adoption in clinical practice including population screening. This research attempted to improve biomarker selection by:

- Understanding the full range of variability in non-neoplastic and/or non-diseased tissues;
- Including diseased but non-neoplastic controls (i.e. colitis/inflamed tissues) in the analyses;
- Introducing a filter to identify possibly neoplasia-specific markers which suggest qualitative versus quantitative change;
- Choosing strong univariate candidates for building multivariate classification models. The advantage of this approach is a potential simplification of future assay development activity.

A complete list of validated gene expression biomarkers is shown in Appendix Table D.17, p. 301.

9.5.1 A list of biomarker candidates

A subset of twenty-two genes was identified by both differential display discovery and microarray analysis to be over-expressed in neoplastic tissues relative to non-neoplastic controls and were likewise validated in the hypothesis testing experiments using a custom microarray (See Section 8.8.1, p.173). Eight of these genes have also been shown in published research to be up-regulated in colorectal neoplasia while the remaining 14 are relatively undescribed in terms of their potential utility as biomarkers for colorectal neoplasia.

These biomarker candidates are compelling for two reasons. First, they have demonstrated differential expression in three independent experiments carried out in the course of this research (and, in some cases, in the work of other scientists). The over expression was observed to be relatively large in terms of both degree and mean difference. Some of the genes also exhibit the "turned-on"

pattern in the discovery data although the validation data were not suitable for testing this hypothesis.

A second reason these candidates are compelling is because of the nature of the data under study. This research represents perhaps the largest known focused study combining discovery and validation data from both adenomatous and cancerous tissues. The use of both normal and non-neoplastic disease RNA extracts in the control group of the discovery data provides further support that the resulting gene patterns are relatively robust for discriminating colorectal neoplasia from non-neoplastic controls.

While this research identifies a surplus of biomarker candidates with high sensitivity and specificity, the table of twenty-two biomarkers identified in Table 8.8, p. 175 provides a useful starting point for future biomarker research. Though the hypothesis of "neoplasia-specific" expression is untested here, the genes shown in Table 8.7,p. 171 which appear to exhibit a "turned-on" gene expression profile also warrant further study.

9.6 Future work

9.6.1 Biomarker assay development

Pepe et al. describe a five-phase pathway that is appropriate for cancer biomarker development [Pepe et al., 2001]. In the context of that framework, this research completes "Phase 1: Preclinical exploratory studies". According to Pepe, the aims of Phase 1 are to a) identify leads for assay development and b) prioritize these leads.

Lead candidates have been prioritised in this research based on redundant discovery in both discovery data sets, performance characteristics for classification and the suggestion of neoplasia-specific gene transcription profiles. In ongoing research outside the scope of this thesis, the author and collaborators have initiated "Phase 2: Clinical assay development" as described by Pepe et al., aimed

at detecting proteins and peptides hypothesized to be differentially translated based on the differential transcription discovered and validated here. Further, in addition to investigating protein-based marker tests, *in vitro* assays should (and will) also explore the utility of both RNA- and DNA-based diagnostic tests.

In addition to over-expressed biomarker candidates, this research identified a large number of under-expressed biomarkers. For convenience and brevity, this thesis has focused particularly on expression markers that are increased in neoplastic tissues. Nevertheless, one could alternatively aim to discover down-regulated markers to the exclusion of over-expression results. One reason I have chosen to focus on over-expressed biomarker candidates is because of the potential theoretical difficulties of measuring all predicted molecules related to a given down-regulated gene expression candidate, as discussed below.

Application of these candidate biomarkers to *in vitro* assays will be extended to all molecularly-related forms of these candidates, including possibly translated protein products. The presence of non-neoplastic cells and molecules in either circulating blood or stool excreta complicates the clinical utility of directly measuring RNA or proteins translated from down-regulated genes. Measuring *the absence* of a signal may be difficult to achieve in diagnostic tests because the relative contribution of cells or molecules from non-neoplastic sources could be much greater than the contribution from neoplastic tumours in a non-invasive specimen. If so, the diagnostic test would involve measurement of relatively small concentration drops between non-diseased and diseased specimens. This assumption of a mixture of neoplastic and non-neoplastic molecules in the clinical specimen is likely to be valid in the case of circulating blood and may also be valid for cell exfoliation from a single neoplastic tumour compared to the exfoliation of the otherwise normal colonic lumen.

Interestingly, there are several suggestions in the literature that the identification of "normal" (i.e. non-neoplastic) molecular markers of colorectal epithelial cells in the peripheral blood could be useful for detection and prediction of colorectal metastasis [Huang et al., 2003, Guadagni et al., 2001]. While these studies provide little evidence of early detection (e.g. adenoma) by measuring biomarkers

of otherwise normal colorectal epithelial cells, the studies support the notion that measuring down-regulated proteins originating from the colorectum will be difficult.

Consequently, future marker research could be specifically targeted toward those candidates that we hypothesise herein are at least up-regulated in neoplasia, and preferably, are neoplasia "specific", i.e. qualitatively changed. In this case, the discriminant rule simplifies to the presence or absence of the target biomarker molecule, where presence of the molecule corresponds to a positive assay result for neoplasia.

Biomarkers down-regulated in neoplasia may still be useful, however. For example, such markers could be measured in assays involving epigenetic changes (i.e. silencing) associated with lowered gene transcription. Hypermethylation is a convenient example. Rather than measuring lower concentration of the biomarker itself, one could possibly measure the *presence* of epigenetic factors (e.g. hypermethylation resulting in down regulated expression) that may be associated with such transcriptional silencing. Several methods have been well established to measure methylation changes including, for example, methylation specific PCR [Rand et al., 2005].

9.6.2 Further research directions

Improved biological understanding

Over the course of this research many candidate biomarkers for colorectal neoplasia were identified and validated based on transcript expression. No attempt has been made to elucidate the biological processes associated with these expression changes for even a single molecule. The goal has been to construct models for phenotype classification that will lead to future assay development research, and ideally, to improved patient outcomes through early disease detection. Nevertheless, improved understanding of the underlying biological changes associated with these neoplastic signatures could enable both better diagnostic tools

and possibly insights related to the neoplastic transformation itself. Improved understanding of neoplasia aetiology could also suggest better therapeutic and prophylactic approaches.

This research suggests a number of potential avenues concerning the molecular biology of gene expression changes in colorectal neoplasia. In particular, the hypothetically neoplasia-specific transcripts (i.e. those that exhibit the prototypical "on" pattern) may provide a convenient, simplified basis for research aimed at such improved molecular understanding. Assuming that some of these genes are indeed switched "on" during the early stages of colorectal neoplasia, the question arises as to the mechanism of such qualitative change. Does a subset of the "turned-on" genes suggest a common denominator, e.g. a common transcription factor, binding motif, etc.?

Improved phenotype-specific gene detection

The notion of applying mathematical algorithms to predict phenotype-specific gene expression patterns introduced in this work has not been previously reported. Nevertheless, the method introduced here is naive, and the method based on an underlying assumption that most genes will not be specifically transcribed in any given cell or tissue specimen of a particular phenotype. Consequently, the majority of genes on a genome-wide microarray should be theoretically "off". When applied to these data, the resulting expression profiles of qualitatively expressed genes generally agree with a prototypical "binary" expression pattern.

Given the utility of identifying and understanding such genes, this method is worthy of further study and development. In particular, more sophisticated estimates of the "off" expression profile would be useful. Initial experiments using variance-based estimates in place of absolute expression level changes to set "on/off" expression thresholds yielded very similar results in these discovery sets (data not shown). Nevertheless, a systematic "discovery" research project aimed at identifying such "on/off" genes, in particular, may be worthwhile.

9.7 In closing

This thesis describes the discovery and validation of biomarker candidates for colorectal neoplasia. The candidates include both previously described and novel candidates including biomarkers which discriminate both adenomatous and cancerous RNA from non-neoplastic controls with a high degree of prediction accuracy. These biomarker leads will be studied for assay development and clinical research aimed at improving health outcomes related to colorectal cancer.

Appendix A

Colorectal gene expression literature

A.0.1 Differential display literature

Early RNA profiling aimed at identifying gene expression differences between two sources of mRNA involved (suppression) subtractive hybridization developed by Lee et al. [1991], an extension of a technique introduced earlier by Davis et al. [1984]. Subtractive hybridization was first used to construct colorectal cancer cDNA libraries by CW et al. [1990].

In 1992 Liang and Pardee developed a PCR-based technique called differential display to amplify cDNA reverse transcribed from mRNA [Liang and Pardee, 1992]. This technique enabled discovery of differentially expressed messenger RNA of interest by comparing PCR products amplifying cDNA synthesized from different mRNA populations. Yeatman and Mao applied differential display to explore colorectal cancer metastasis to the liver in 1995 [Yeatman and Mao, 1995]. One of the discovery arms explored in this thesis is based on differential display technology.

Also in 1995, Victor Velculescu developed the serial analysis of gene expression (SAGE) technique [Velculescu et al., 1995] involving generating libraries of ex-

pressed sequence tags for comparison between phenotypes. Whereas differential display techniques generally involve observing phenotypic band differences using gel electrophoresis, SAGE involves computer intensive analysis of automated sequencing data from concatenated sequence tags. This technique was applied by Zhang et al. in 1997 to discover 500 (out of approximately 300,000) differentially expressed transcripts in neoplastic colon cells compared with normal controls [Zhang et al., 1997]

A.0.2 Microarray-based discovery

There are now numerous reports in the literature involving microarray-based discovery of colorectal neoplasia markers, including both cDNA microarrays and synthetic oligonucleotide arrays. To put this expansion in to perspective an *ad hoc* analysis of the magnitude of this growth was carried out by simply counting the number of PUBMED (`http:\\www.pubmed.org`) returns by year for a query using the search term: "gene expression colorectal". The results of this simple experiment are shown in Figure 1.1. Given the number of papers in this field, a

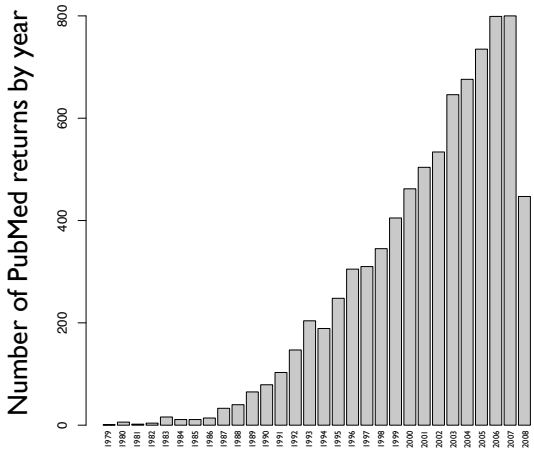


Figure 1.1:

complete analysis of all research is not practical. Consequently, a survey of key reports is presented here as Table A.1.

Table A.1: Survey of microarray experiments measuring gene expression in the colorectum

Year	**Author**	**Platform**	p **genes**	**Sample**	N **(dis/norm)**	**Ref**
1991	Augenlicht et al.	cDNA	30	tissue	19/6	[Augenlicht et al., 1991]
1999	Alon et al.	Affy Hum6000	6800	tissue	40/20	[Alon et al., 1999]
1999	, Backert et al.	cDNA (custom)	588	tissue & cells	12/12	[Backert et al., 1999]
2001	Notterman et al.	Affy (HU6500/HU6800)	6,600	tissue	22/22	[Notterman et al., 2001]
2001	Yang et al	cDNA (custom)	8,063	tissue	3	[Yang et al., 2001]
2001	Buckhaults et al.	SAGE	21,343	tissue	4/2	[Buckhaults et al., 2001]
2001	Giordano et al.	Affy (HuGeneFl)	7,129	tissue	51	[Giordano et al., 2001]
2001	Ramaswamy et al.	Affy (Hu6800&35KsubA)	16,063	tissue	11/11	[Ramaswamy et al., 2001]
2001	Takemasa et al.	cDNA (custom)	4600	tissue	12/12	[Takemasa et al., 2001]
2001	Kitahara	cDNA (custom)	9,216	tissue	20/20	[Kitahara et al., 2001]
2001	Lin et al.	cDNA (custom)	9,216	tissue & cells	12/12	[Lin et al., 2001]
2001	Lechner et al.	cDNA (2 Atlas)	316	tissue & cells	0/1	[Lechner et al., 2001]
2001	Miwa et al.	cDNA (custom)	9,216	tissue	16/16	[Miwa et al., 2000]
2001	Su et al.	Affy (U95a)	12,000	tissue	10/0 (multi-class)	[Su et al., 2001]
2002	Platzer et al.	Affy (custom)	55,000	tissue	23/0	[Platzer et al., 2002]
2002	Agrawal et al.	Affy (Hu6800&Hu95A)	12,000	tissue	60/10	[Agrawal et al., 2002]

Continued on Next Page...

Table A.1 – Continued

Year	Author	Platform	p genes	Sample	N (dis/norm)	Ref
2002	Birkenkamp et al.	Affy (HuGeneFl&35KsubA-D	42,843	tissue	40/10	[Birkenkamp-Demtroder et al., 2002]
2002	Gerritsen et al.	Affy (Hu6800)	7,129	tissue & cells	6/6	[Gerritsen et al., 2002]
2002	Dieckgraefe et al.	Affy (Hum 6000)	23,040	tissue	8/3 (U. colitis)	[Dieckgraefe et al., 2002]
2002	Lin et al.	cDNA (custom)	23,040	tissue	20/0	[Lin et al., 2002]
2002	Zou et al.	cDNA (custom)	8,000	tissue	13/13	[Zou et al., 2002]
2002	Ichikawa	cDNA (custom)	20,784	tissue	7/16/12 (Ad/Ca/Mets)	[Ichikawa et al., 2002]
2003	Masuda et al.	Affy (Hu95A)	12,000	tissue	7/3 (U. colitis)	[Masuda et al., 2003]
2003	Muro et al.	adaptor tagged RT-PCR	1,536	tissue	100/11	[Muro et al., 2003]
2003	Brunschwig et al.	Affy (custom)	–	tissue	–	[Brunschwig et al., 2003]
2003	Frederiksen et al.	Affy (HuGeneFL)	6,800	tissue	20/5	[Frederiksen et al., 2003]
2003	Kemmner et al.	Affy (HU95A)	12,000	Pooled pat. samples	4/4	[Kemmner et al., 2003]
2003	Wiliams et al.	cDNA (custom)	9,592	tissue	20/20	[Williams et al., 2003]
2003	Neumann et al.	cDNA (2 Atlas)	199/597	tissue	1	[Neumann et al., 2003]
2003	Buckhaults et al.	SAGE	21,321	tissue	20/0	[Buckhaults et al., 2003]
2003	Mori et al.	cDNA (LLNL)	8,064	tissue	41/0	[Mori et al., 2003]
2003	Kim et al.	Oligo (β-catenin mutn only)	121	tissue & cells	74/0	[Kim et al., 2003a]

Continued on Next Page...

Table A.1 – Continued

Year	Author	Platform	p genes	Sample	N (dis/norm)	Ref
2003	Glebov et al.	cDNA (multiple)	6,500/9,000	tissue	0/50	[Glebov et al., 2003]
2003	Clarke et al.	cDNA (LLNL)	4,132	tissue	18/0	[Clarke et al., 2003]
2003	Jubb et al.	Affy (Hu133A)	2	tissue	222/211	[Jubb et al., 2004]
2004	Crott et al.	Affy (U34)	9,000	(rats)	–	[Crott et al., 2004]
2004	Friedman et al.	Affy (custom)	NA?	tissue & cells	85/28	[Friedman et al., 2004]
2004	Kusakai et al.	cDNA (BD cancer array)	241	tissue	56/15	[Kusakai et al., 2004]
2004	McIver et al.	cDNA (custom)	8,000	tissue	68/68	[McIver et al., 2004]
2004	Mori et al.	cDNA (custom)	8,064	tissue & cells	85/26	[Mori et al., 2004]
2004	Mennerich et al.	Affy (Hu133A&B,cancer)	35,000 / 241	tissue	58/58	[Mennerich et al., 2004]
2004	Kwon	cDNA	4,080	tissue	12/12	[Kwon et al., 2004]
2004	Croner et al.	Affy (HU95A)	12,625	tissue	1	[Croner et al., 2004]
2004	Koehler	cDNA (Atlas)	–	tissue	25/25/14 (Ca/N/Met)	[Koehler et al., 2004]
2004	Wang et al.	Affy (Hu133A)	35,000	tissue	74	[Wang et al., 2004]
2004	Bertucci	cDNA (custom)	8,000	tissue	50/50	[Bertucci et al., 2004]
2005	Mori	cDNA Atlas Glas	8,000	tissue	6/6	[Mori et al., 2005]
2005	Chiu	cDNA (ABC)	8,000	tissue	4/4	[Chiu et al., 2005]
2005	Sugiyama	cDNA (HumCanPath)	96	tissue	11/11	[Sugiyama et al., 2005]
2005	D'Arrigo	cDNA (custom)	7,864	tissue	10/10 (mets)	[D'Arrigo et al., 2005]

Continued on Next Page...

Table A.1 – Continued

Year	Author	Platform	p genes	Sample	N (dis/norm)	Ref
2005	Salahshor	cDNA (custom)	19,200	tissue	3/1 (Ad/Norm)	[Salahshor et al., 2005]
2005	Eschrich	cDNA (TIGR)	31,872	tissue	78/0 (prognosis)	[Eschrich et al., 2005]
2006	Jansová	cDNA Hum19K	19,000	tissue	9/9 (mets)	[Jansova et al., 2006]
2006	Ohmachi	cDNA Agilent	12,814	tissue	16/15	[Ohmachi et al., 2006]
2006	Galamb	cDNA Atlas glass	7,864	tissue	10/6/6 (Ad/Ca/IBD)	[Galamb et al., 2006]
2006	Chowdary	Affy U133A	22,215ps	tissue	42/0	[Chowdary et al., 2006]
2007	Habermann	cDNA (custom)	9,000	tissue	16/17/20/13 (Norm/Ad/Ca/Mx)	[Habermann et al., 2007]
2007	Grade	Oligo (NCI)	21,543	tissue	30/30	[Grade et al., 2007]
2007	Sabates-Bellver	Affy U133Plus2	55K	tissue	32/32	[Sabates-Bellver et al., 2007]
2007	Ojima	Agilent Hu25K	24,479	tissues/cells	30 (all tumour)	[Ojima et al., 2007]
2007	Wiese	Affy Hu95A/Av2	–	tissues/cells	29 tumour+2 lcm	[Wiese et al., 2007]
2007	Hong	Affy U133Plus2	55K	tissue	12/10	[Hong et al., 2007]

Continued on Next Page...

Table A.1 – Continued

Year	Author	Platform	p genes	Sample	N (dis/norm)	Ref
2007	Yajima	Affy U133A	22,215ps	fecal colonocyte	23/15	[Yajima et al., 2007]
2007	Maglietta	Affy U133A	22,215ps	tissue	50/47	[Maglietta et al., 2007]
2007	Ayers	Affy U133A	22,215ps	tissue	118/0	[Ayers et al., 2007]
2007	Watanabe	Affy U133Plus2	55K	tissue	10 UC-Neop/43 UC-NonNeop	[Watanabe et al., 2007]
2007	Collado	cDNA (Res. Genetics)	15,552	tissue	12/8	[Collado et al., 2007]
2008	Galamb	Affy U133Plus2	55K	tissues/blood	22/20/11/21/11 (Ca/Ad/Hp/IBD/Norm)	[Galamb et al., 2008]
2008	Kim	Olgo HumCancer3K	3,096	tissues (serrated ad)	5/5	[Kim et al., 2008b]
2008	Han	Affy U133Plus2	55K	blood	16/15	[Han et al., 2008]
2008	Kim JC	cDNA (custom)	21,000	tissue	84/84	[Kim et al., 2008a]
2008	Aerssens	Affy 133Plus2	55K	tissue	36 IBD/25	[Aerssens et al., 2008]
2008	Yu	Affy 133A	21,000	tissue	9/9	[Yu et al., 2008]

Following the gene expression work of Golub et al. [1999] in leukaemia, Alon et al. [1999] examined 62 colorectal tissues using the Affymetrix Hum6000 array which contained probes on four separate chips for approximately 3,200 full length human genes and 3,200 EST's taken from the Human Genome Project. To analyse those data, Alon used a hierarchical clustering algorithm based on binary trees to cluster the tissues and the genes. The results identified two tissue clusters that the authors attribute to correspondence with the overall cell composition of tissue biopsy. According to the authors, "It is expected that the normal tissue samples include a mixture of tissue types, while the tumour samples are biased to epithelial tissue of the carcinoma." Observing that five of the top 20 most differentially expressed tissues were muscle related genes, Alon supports his mixed cell type hypothesis by calculating a "muscle index" based on 17 EST sequences with homology to smooth muscle genes. Using this index, Alon observed that while normal tissue demonstrated a high muscle index, the tumour tissues were found to have a relatively lower index. Furthermore, outlier normal tissues that were "mis-clustered" with tumour samples were shown to have relatively low muscle index and vice versa leading the authors to conclude that such outliers could be accounted for by tissue composition [Alon et al., 1999]. Interestingly, a similar phenomenon was observed during this research when analysing another publicly available data set and similar conclusion was reached.

In a follow-up study from the same research team, Notterman et al. [2001] used the Affymetrix HU6500 GeneChip to compare expression between 18 colon cancer and matched normal specimens and the HU6800 GeneChip to compare 4 colorectal adenomas with matched normal tissues [Notterman et al., 2001]. The authors used univariate statistical tests (Student's t test or Mann-Whitney U test) to explore gene expression variation between the tissue classes. However, a hierarchical clustering algorithm was also used to analyse the global gene expression changes between the tissues using a subset of 1,096 genes (to handle differences in the two chip platforms.) By visual inspection the authors identified three broad tissues clusters corresponding approximately to adenoma, cancer

and normal tissues. A number of cell cycle regulators, oncogenes, etc. were identified in the two disease sets that are worthy of follow-up. Interestingly, genes related to smooth muscle and connective tissue were over-expressed in normal tissues relative to cancer tissues, similar to the findings of Alon.

Both Alon et al. and Notterman et al. employ hierarchical clustering techniques to visualize and explore the gene expression profiles of the sample tissues. This unsupervised clustering technique sorts individual genes or tissues according to a two-way pair wise average linkage classifier so that individuals with similar scores (of the chosen metric) are near each other on the graph. This method is similar to the phylogenetic trees used in comparing evolutionary lineage. Perhaps because of the precedent set by this early work, many researchers also employ clustering techniques as the primary analytical method. Further, both of these authors explicitly identify those genes differentially expressed in a univariate sense between the tissue classes of interest. There is little attention paid to high-dimensional gene expression relationships within the data. Rather, over-expressed and under-expressed genes are tabulated and weighed for potential relevance in isolation without regard to the inter-dependent (network) nature of gene concentrations.

Studies by Yang et al. [2001] and Clarke et al. [2003] measured the effects on gene expression of patients undergoing treatment with sulindac and 5-fluorouracil (5-FU), respectively. These studies are relatively unique in their aim of measuring drug interactions at the gene expression level in live human patients and may represent the first clinical studies in colorectal cancer to profile gene expression in response to chemical prophylaxis and chemotherapy, respectively [Yang et al., 2001, Clarke et al., 2003].

Yang et al. measured pooled rectal biopsy specimens from three patients at increased risk of cancer before and after a one month treatment of 300 mg sulindac/day. Sulindac, a non-steroidal anti-inflammatory drug (NSAID), has been shown to mitigate intestinal tumours in FAP patients and to inhibit tumour formation in the MIN mouse model [Giardiello et al., 1993]. Among the interesting findings, the authors observed decreased expression of seven genes of the

immunoglobulin family and increased expression of the cyclin dependant kinase inhibitor, p21WAF1/cip1. The lowered expression of immune-related genes is presumed to be a natural consequence of the anti-inflammatory drug lowering the number of lymphocytes in the biopsy specimen. The authors further investigate the role of p21 by creating p21WAF1/cip1 knockout mice to show that p21 was, in fact, required for sulindac activity in APC+/- mice. Though the authors recognise that a number of confounding variables could influence gene expression (e.g. diet, genetic background, etc.), their discovery of key genes shown to be involved with tumour progression and drug action (p21) provide evidence to the value of this experimental design.

Clarke et al. also studied drug effects on gene expression in rectal cancer but this study looked at the chemotherapeutic effects of a combination treatment of 5- fluorouracil (5-FU) and mitomycin (MMC) in patients with advanced disease. All 18 patients in this study were diagnosed with T3 or T4 rectal cancer and each had a significant risk of incomplete surgical clearance. To better understand the molecular pharmacology of cancer, the authors measured gene expression in biopsy specimens taken prior to, and during, a course of preoperative chemoradiotherapy. In the baseline analysis of tumour specimens to normal mucosa, the authors observed a higher level of expression of gene families typically associated with a mixed cell composition. The identified genes families include colonocyte genes, hematopoietic and immunoglobulin genes, and smooth muscle genes in the normal specimens. This observation agrees with the findings of Alon and Notterman discussed above. The authors also reported an over expression of *MYC* in tumour tissues prior to treatment and a corresponding decrease in *MYC* gene expression in the post-treatment tumour biopsy samples. This observation led the authors to conclude that decreased *MYC* expression or activity could participate in the anti-tumour mechanisms of MMC/5-FU treatment.

Buckhaults et al. [2001] of the Kinzler-Vogelstein laboratory used a large SAGE library (290,394 tags for 21,343 transcripts) to measure transcription differences between normal tissues, colorectal adenomas, and cancers. Of the nine transcripts they identified to be at least 20-fold over expressed in cancers and ade-

nomas relative to normals, six transcripts were predicted to be either secreted or cell-surface expressed. The genes include *TGFβI*, *LYS*, *RDP*, *MIC-1*, *REGA* and *DEHL*; the results were confirmed by RT-PCR in epithelial cells extracted from the tumour tissue by immunopurification.

Several publications report the use of gene expression to characterize and classify tumour samples from among multiple tumour tissue types. Giordano et al. [2001] measured gene expression in 154 primary adenocarcinomas from the lung, colon and ovary, Ramaswamy et al. [2001] examined 218 tumour samples comprised of 14 tumour types and Su et al. [2001] measured 175 tissues from 10 tumour classes. The primary aim of each of these studies was to differentiate tumour samples based on gene expression. Interestingly, these groups were also among the first (all three published in October 2001) to apply relatively strong supervised machine learning techniques (k-nearest neighbour and two support vector machines, respectively) to discriminate the multi-class data. In a later study, Buckhaults et al. [2003] used SAGE to analyse 62 tumour samples taken from ovarian, breast, colon, and pancreatic adenocarcinomas for the purpose of identifying the primary tumour location from a secondary metastasis. Buckhaults et al. used both a self-organising map (unsupervised) algorithm and a modified support vector machine (supervised) algorithm to analyse their high dimensional data.

Takemasa et al. [2001] is of particular relevance to this thesis because the authors appear to have utilized a similar strategy to that of this thesis for mining the transcriptome by combining a "discovery"-based method with hypothesis driven gene selection. To do this, Takemasa et al. constructed a specialised "Colonchip" by spotting 4,608 separate clones that were isolated from a 30,000 clone library derived from late stage colorectal cancer, matched normal tissues, and liver metastatic cancers. The authors also included 170 "conventional" genes suspected to be involved in colorectal carcinogenesis on the custom chip. By analysing an additional set of 12 colon and 12 normal samples with dual-labelled (Cy5/Cy3) cDNA targets, the authors identified 59 genes (23 up in tumours, 36 down in tumours) that were differentially regulated by two-fold

or greater. Multivariate techniques to explore multi-gene interactions were not used.

Platzer et al. [2002] constructed a massive, 55,000 transcript microarray using Affymetrix's oligonucleotide system that contained all of the known human genes in the public domain at the time of design. Interestingly, this chip size is roughly equivalent to the eventual size of the Affymetrix U133plus2 system. The authors used these chips to compare chromosomal amplification with gene expression profiles in 15 colorectal cancer specimens and 8 colon cancer liver metastases. This study focused specifically on transcripts that were judged to map from four chromosomal locations found to be commonly amplified in colon cancers (7p, 8q, 13q, and 20q.) Of the 2,146 transcripts originating from within these regions, only 81 (3.8%) were discovered to demonstrate at least 2 fold increase in expression. Based on this work, the authors conclude that while chromosomal amplifications may be common in colon cancer, increased expression of transcripts from such regions is relatively rare. This finding is intriguing and perhaps slightly controversial given the strong evidence of frequent chromosomal instability in colorectal cancer (see earlier discussion of the CIN pathway). Regardless, this work is worthy of follow-up to better clarify the relationship between aneuploidy and gene expression.

Surprisingly, the authors did not comment on gene expression for probes outside of the four "amplified" chromosomal regions, despite the fact that a custom "total" genome chip was created.

A unique and elegant marker selection approach was demonstrated by Gerritsen et al. [2002] by combining gene expression data from in vitro models with in vivo data using sophisticated bioinformatics techniques. Working from a conceptual hypothesis that angiogenesis markers of interest in colon tumours should be of stromal (i.e. not epithelial) origin, the authors analytically subtracted (*in silico*) genes over-expressed in colon cancer cell culture from a super-set of candidate markers derived by intersecting established angiogenesis genes with a database of colon tumour genes. The authors report a resulting list of 24 candidate endothelial-derived angiogenesis associated genes that may be of utility on the

custom oligonucleotide chip constructed in this thesis.

Given the massive accumulation of biological data (and in some cases, knowledge) being gathered within public databases (e.g. NCBI web portal `www.ncbi.nlm.nih.gov`), I suggest that Gerritsen's technique is under-utilised and the potential of this approach should be further explored. Nevertheless, I find interesting that in an editorial of the same issue of Gerritsen's publication, editors Aird et al. [2002] provide a flawed perspective about the value of this work, in my opinion. The editors refer several times to the "overwhelming number of genes" differentially expressed between the tissues of interest. Consequently, out of concern for generating "false positive" results based on too many genes, they support the use of Gerritsen's approach to filter the number of genes to analyse. I agree that Gerritsen's approach is valuable. However I find the innovative contribution to be how that group used in silico mining techniques to refine the data analysis not by reducing the data, but by increasing the information content of the experiment.

While most gene expression research related to colorectal cancer is focused on late stage cancer and metastasis, this thesis attempts to identify molecular markers useful for diagnosing precancerous colorectal adenomas. In fact the earliest known example of gene expression analysis of colorectal adenoma tissues relative to non-neoplastic controls is described in this thesis based on the unpublished work of James and Kazenwadel [2002].

The first example of gene expression analysis using adenoma tissues is presented by Lin et al. [2002] who used a custom cDNA array built using 23,040 sequences taken from the NCBI's UniGene database to analyse 11 colorectal cancer and 9 colorectal adenoma tissues vs. matched normal specimens. Based on a relatively weak differential display criteria ($>$ 2 fold change in at least 50% of the tissues), the authors found 427 genes differentially expressed (51 up, 376 down). Using a two-dimensional hierarchical clustering algorithm with 771 genes, the authors were also able to distinctly cluster the adenoma and carcinoma samples. By using the normal colonic tissue as the second (Cy5) label in the two colour (Cy3/Cy5) cDNA hybridisation experiment, the authors analysed the data for

just two classes: adenoma vs. cancer. While the clustering techniques used by Lin et al., provide a modest degree of relationship information about gene expression between the two tissue classes, these data represent a missed opportunity to use sufficiently strong high-dimensional analytical techniques for discovering markers differentiating adenomas and cancers from normal specimens.

Since the work of Lin et al., few other researchers have explored differential gene expression in colorectal adenomatous tissues. Ichikawa et al. [2002] measured 7 adenoma tissues versus 16 cancer tissues using a custom cDNA to assemble what the authors describe as a predictor of malignant phenotype. The predictor involving 335 genes diagnosed 12 additional specimens (5 cancer with metastases, 7 metastatic tissues (liver and lung)) correctly as cancer. The predictor, however, also identified three of the original seven adenomas as cancerous. It is surprising that the authors did not include non-neoplastic test tissues in this study.

Galamb et al. [2006] analysed 10 adenomas and 6 cancers and 6 inflammatory bowel disease (IBD) specimens using Atlas Glass 1K cDNA microarrays. While the content of the microarray chip is limited with only 1,081 gene targets, the inclusion of IBD specimens in this study is unique and notable in the literature. Recently, Galamb et al. followed this initial research with a much larger experiment. Using full-genome Affymetrix HGU133plus2 (55,000 probesets) Galamb et al. again analysed a wide range of specimens including cancer, adenoma, hyperplastic polyps, IBD and healthy normal controls [Galamb et al., 2008]. For comparison, RNA extracts from 30 peripheral blood samples (19 cancer, 11 healthy controls) were included. Using sophisticated modern data analysis techniques (e.g. gcRMA normalisation, significance analysis of microarray, bootstrap error prediction, random forest classification, etc.) Galamb et al. were able to discriminate most phenotypes from each other. In particular Galamb et al. identified *KIAA1199*, *FOXQ1*, and *CA7* to be differentially expressed in colorectal adenomas relative to normals.

Habermann et al. [2007] used a 9K cDNA microarray platform to measure the complete adenoma-carcinoma sequence using 33 specimens, including 3 normal tissues, 8 adenomas, 15 primary sporadic cancers and 7 metastatic liver tissues.

The authors identified 58 genes differentially expressed between adenoma and normal tissues (20 up, 38 down); 116 genes differentially expressed between cancer and adenoma (80 up, 36 down); and 158 genes differentially expressed in liver metastases and cancer tissues (138 up, 20 down). The observation of more genes down-regulated in neoplastic adenomas relative to normal controls while there are more genes with higher expression with increasing disease state is in agreement with the literature and the results of this thesis. Although this research appears to be of a high standard, the combination of a small microarray and relatively limited number of specimens suggests caution with respect to these data.

The largest study aimed at measuring adenoma gene expression profiles is by Sabates-Bellver et al. [2007]. By measuring 32 prospectively collected adenoma tissue samples and an equal number of normal controls using the Affymetrix U133Plus2 microarray, Sabates-Bellver et al. discovered over 15,000 probesets to be differentially expressed in adenomas, representing more than 25% of the probesets available on the microarray. The authors also concluded that *KIAA1199* could be a novel marker of colorectal neoplasia in agreement with the results of Galamb et al. [2008].

Recently, Kim et al. [2008b] became the first group to apply gene expression microarrays to serrated adenomas. Kim et al. applied a custom cDNA microarray to five serrated adenoma tissues and matched normal controls. The authors identified 73 genes up-regulated by 2-fold in serrated adenomas and 51 genes down-regulated in normal mucosal specimens. In particular, the authors identified *TNFRSF10A* to be over-expressed and *BENE* and *RARA* to be down-regulated in serrated adenomas. Results for these three genes (only) were validated by RT-PCR.

Only one publication in the identified literature reports the use of gene expression data in colorectal samples to predict survivability. Muro et al. [2003] measured gene expression in 100 colorectal cancer samples and 11 normal samples using adaptor- tagged competitor PCR for 1,536 genes of interest. In addition to discovering an expression pattern between the multiple target classes (normal,

tumour, distant metastasis) the authors also analysed the capacity of the tumour classifier gene set (12 genes) to predict survivability with significant results.

Several groups have reported the use of microarray data to classify colorectal tumours by stage (Dukes', Astler-Coller modified Dukes, or TNM). Agrawal et al. [2002], Birkenkamp-Demtroder et al. [2002], Frederiksen et al. [2003], and Wang et al. [2004] use Affymetrix GeneChip arrays to identify genes differentially expressed between the tumour stages.

Kemmner et al. [2003] used a 12,000 probe Affymetrix gene chip to measure 39 glycosyltransferases and 10 sulfotransferases in pooled samples of colonic epithelium extracted by laser micro-dissection. This research appears to be unique as five samples of healthy, normal mucosa (i.e. taken from disease-free individuals) were compared to two classes of colorectal cancer specimens stratified by low or high risk of tumour-dependent death.

Mori et al. [2004] examined 85 primary colon cancers, 26 normal colonic mucosal samples and colon cancer cell lines using an in house 8,064 sequence cDNA array. The cancer tissues and cell lines were classified by microsatellite instability status as MSI-High or non-MSI-High. The authors used univariate analysis techniques to find significant under-expression of 81 (of 8064) genes in MSH-H samples relative to non-MSI-H. These under-expressed genes were then searched for CpG sites using public databases (e.g. NCBI) to yield 46 potential targets of hypermethylation-mediated gene silencing. This is a novel-use of gene expression data for MSI-H colorectal cancers in the literature, although differentially expressed genes between the tissue classes are not discussed.

Several well described molecular mediators of colorectal lesion formation have been shown to be targets of mutations, e.g. APC, β-catenin, k-Ras and others [Bodmer et al., 1989, Nishisho et al., 1991, Vogelstein et al., 1988, Fearon and Vogelstein, 1990]. Kim et al. [2003a] applied specific knowledge of β-catenin mutations to create a custom oligonucleotide array with hybridisation specificity for 110 specific β-catenin mutations. By analysing 74 colorectal carcinoma specimens and 31 colorectal cancer cell lines, the authors observed that the fre-

quency of β-catenin mutations was higher in both MSI tumours and cancers of the proximal colon. This work has potential relevance to this thesis by applying the general principal that one need not be limited by the "public domain" of commercially available microarrays. It is possible that this research could be successfully extended to any molecular targets that exhibit a high number of mutational hotspots or polymorphisms.

Another significant aspect of this thesis will include an analysis of gene expression variation across a normal colon. This analysis is vital to the primary goal of identifying molecular biomarkers for colorectal cancer in order to avoid anatomical bias in presumed biomarkers. There is a general consensus that proximal and distal tumours have broadly distinct molecular pathogenesis aetiologies and that these differences stem from subtle tissue differences across the colon (reviewed in Iacopetta [2002]). One objective of this thesis, will therefore be to explore the anatomical variation in the colon across five discrete segments from the caecum to the rectum. In a 2003 publication, Glebov et al. [2003] provide proof of concept for this approach by simply comparing the gene expression variation between the left and right colon. Using three different cDNA microarray platforms Glebov et al. measured gene expression from standard pinch biopsies of both the ascending and descending colon in 50 patients. The authors used univariate t tests to identify genes differentially expressed between the two tissue sets and then used a classification algorithm (compound covariate predictor) to use only those differential genes to classify each tissues. This study found that over a thousand genes (1,349) exhibit variation between the ascending and descending colon and their classifier algorithm was able to correctly predict the source of 98/100 samples. Finally, the researchers also measured gene expression in 13 paired samples from foetal colons to find 87 genes with differential left and right colon expression.

Croner et al. [2004] examined one of the central issues of gene expression measurements, tissue preparation. Using the Affymetrix HU95A GeneChip (12,000 probe sets), the researchers compared three alternatives to laser capture microscopy: (1) cryotomy after manual dissection, (2) microscopically assisted

manual dissection, and (3) tumour-cell isolation with Ber-EP4 antibody coated Dynabeads. Based on their analysis of a split RNA sample taken from a single patient, the authors conclude that all three methods are suitable for gene expression experiments but that expression comparisons across different methods should be regarded critically. Surprising, the authors do not include a sample processed by laser capture microscopy which is the more conventional, though costly, approach to obtaining purified samples [Rubin, 2001, Kitahara et al., 2001] Consequently, the question still remains as the effectiveness of these alternatives to the sophisticated laser micro-dissection technique.

The question addressed by Croner, however, is of prime importance to the analysis of gene expression in colorectal adenomas. A review of these studies shows that several groups observe gene expression variation between normal mucosa and colorectal tumours that is attributed to variation in the cellular composition of the biopsy specimens. A definitive gene expression study has not been carried out that explores this hypothesis. Further, the more fundamental clinical vs. biological question that remains is to what degree cell composition (i.e. the ratio of epithelial to non-epithelial cells in a sample) is important when biopsying a clinical specimen designated as normal or diseased. For example, a lower percentage of epithelial cells for a given mass of normal mucosa may represent a mechanical sampling difficulty of "flat" regular mucosa, perhaps regarded as "contamination". On the other hand, the inclusion of non-epithelial cells in a normal specimen may provide further clues about gene expression patterns of the stroma. The differences in tissue morphology, etc. collected in biopsy specimens may provide strong insights to the hunt for diagnostic markers.

Further, upon review of these gene expression measurements in human tissue data, one interesting observation to note is the relative number of over-expressed and under-expressed genes between normal and diseased tissues. Table A.2 outlines these differences for those studies where the data is provided:

With one exception (Bertucci et al.), each study that measures a gene-by-gene comparison between tumour and normal samples finds a higher number of genes under-expressed in tumours compared to normal tissues. Though this trend does

Table A.2: Comparison of the (p) genes over- and under expressed in tumours relative to normal tissues

Study	Cutoff	Over	Under	Reference
Backert	–	2	8	[Backert et al., 1999]
Notterman	4-fold	19	88	[Notterman et al., 2001]
Buckhaults	2-fold	50	192	[Buckhaults et al., 2001]
Takemasa	2-fold	23	36	[Takemasa et al., 2001]
Lin	2-fold	51	376	[Lin et al., 2002]
Kitahara	–	44	191	[Kitahara et al., 2001]
Birkenkamp	abs	27	72	[Bkamp-Demtroder et al., 2002]
Williams	2-fold	574	2058	[Williams et al., 2003]
Bertucci	2-fold	130	115	[Bertucci et al., 2004]

not receive mention in the literature, Birkenkamp-Demtroder et al. observe that based on their study, many of the genes repressed in tumours appear to code for mitochondrial proteins. The authors further hypothesise that decreased RNA transcription could be due to hypermethylation [Birkenkamp-Demtroder et al., 2002]. Several groups have also commented that "normal" colonic mucosa samples appear to be more heterogeneous in nature than colonic adenocarcinoma [Alon et al., 1999, Notterman et al., 2001, Clarke et al., 2003] Obviously, few substantial conclusions can be drawn from this observation, however the relatively higher number of under-expressed genes in tumours compared to normal tissues seems worthy of investigation. Further, while Alon's hypothesis that normal tissues may exhibit a greater degree of mixed cell composition could be related to this phenomenon, one can not exclude the possibility that this lower gene expression in tumours demonstrates some significant, fundamental pathogenic process.

A.1 Conclusion

In conclusion, there are few studies that explore gene expression patterns associated with colorectal adenomas. Among these selected publications, however, there is interesting overlap for particularly differentially expressed genes, such as *KIAA1199*. Of the literature reviewed here, only the work of Galamb et al.

includes both healthy normal controls and non-neoplastic diseased controls (in this case, IBD) for comparison to neoplastic specimens. Given the likelihood that other diseases, including colitis, could affect gene expression patterns in colorectal tissues, this is a key weakness of the prior literature.

Appendix B

Quality control methods

B.1 Aim

The aim of this Appendix is to describe the quality control (QC) methods applied by the author in relation to 548 Affymetrix HG133A & HG133B oligonucleotide arrays resulting in the final selection of 454 GeneChips used for microarray discovery described in Chapter 7.

AUTHOR'S NOTE: The material in this appendix forms internal CSIRO Technical Report 05/205. This work is unpublished.

B.2 Description of Gene Logic data

Gene expression and clinical descriptions for 548 colorectal tissue specimens were purchased from Gene Logic (Gaithersburg, MD, USA) to identify biomarkers for specific colorectal tissue phenotypes and to better understand colorectal biology. The Gene Logic data set was chosen in 2004 after a comprehensive review of public and private data source options.

For each of 548 tissues, the following data were received:

- Raw .CEL files produced by the Affymetrix (Santa Clara, CA, USA) Gene Chip ®microarray system described in Lipshutz et al. [1999],
- Results from HG133A and HG133B chips, a total of 44,928 probesets and
- 81 experimental and clinical descriptors for each tissue.

B.3 Quality control of Affymetrix Gene Chips

Measuring tissue gene expression using high dimensional microarrays involves complex clinical and laboratory processing. The first step in analysing a set of expression arrays, therefore, should be a careful assessment of the data quality to identify and, if appropriate, remove, potentially contaminating arrays from the analysis. This assessment includes basic editing and data review that is fundamental to any multivariate analysis [Chatfield and Collins, 1981].

Affymetrix data quality manuals recommend to focus on five data aspects for quality controlling batches of hybridised Gene Chips [Affymetrix, 2004a]:

1. Absolute chip background (taken to be the lowest 2% of probe intensities)
2. Scale factors used to transform each probeset to an absolute intensity of 100
3. Percentage of probesets (genes) called present
4. Ratio of 3' to 5' binding for housekeeping genes
5. Response of spike-in controls

To assess these QC parameters, the complete set of 548 chips were analysed using 'simpleaffy' [Wilson and Miller, 2005] and 'affy' [Gautier et al., 2004] BioConductor packages that provide convenient access to the Affymetrix QC metrics and normalisation algorithms. BioConductor is an open source R framework that provides a wide range of bioinformatics tools for analysing molecular biological data [Gentleman et al., 2004, R Development Core Team, 2008].

Gene expression levels were calculated by both Microarray Suite (MAS) 5.0 (Affymetrix) and the Robust Multichip Average (RMA) normalisation techniques [Affymetrix, 2001, Irizarry et al., 2003, Hubbell et al., 2002]. The data

were processed as both a single aggregated set of 44K probesets as well as by splitting the data into two subsets, the HG133A chip (22K probesets) and HG133B chip (22K probesets). The availability of two independently hybridised arrays for each tissue sample (Chip A and Chip B) provides a useful means to assess QC parameters in the Gene Logic data set. While the same hybridisation solution for a given tissue will be reacted with both chips, anomalous or outlier results at the tissue-hyb-solution level can be easily observed by inspection, as described below.

B.3.1 Scaling factors

By default, the MAS5.0 normalisation algorithm sets the trimmed mean intensity of every array to an arbitrary level (target=100). The scaling factor is a measure of the scaling applied to each individual array to bring the average intensity to this value.

Figure 2.1 (left) shows the scaling factors for all arrays plotted for Chip A vs Chip B and Figure 2.1 (right) shows the scaling values for A only. These data suggest that the scaling factor applied to Tissue 12204 is exceptionally high for Chip A and on the high range for Chip B; this tissue was scrubbed from the data.

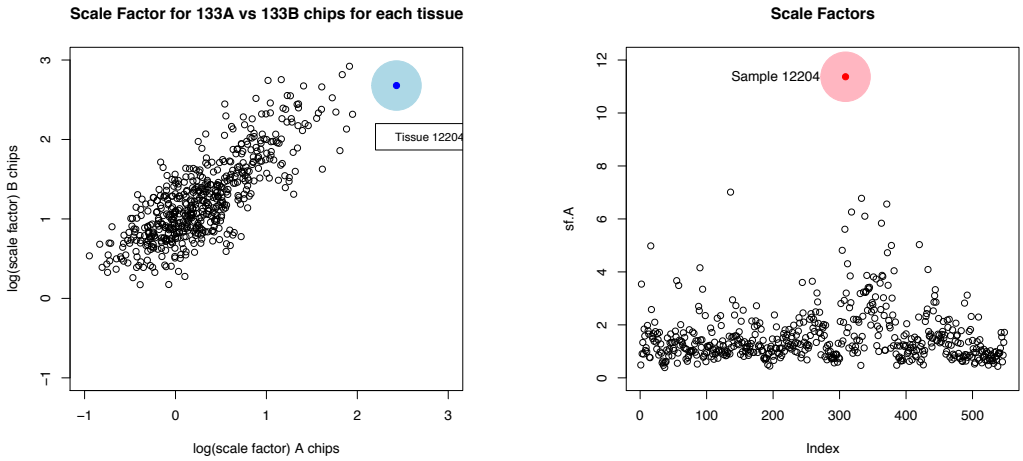


Figure 2.1: scale factors)

B.3.2 Background values

According to Affymetrix guidelines, the background level should be similar across all chips [Affymetrix, 2004a]. Aberrant, high background levels for a particular array may indicate a problem with cRNA concentration, poor washing after hybridisation, or some other experimental anomaly.

Figure 2.2 shows the background values for Chip A vs. Chip B for all arrays. Tissue 3424 clearly has exceptionally high background levels and so was scrubbed from the data.

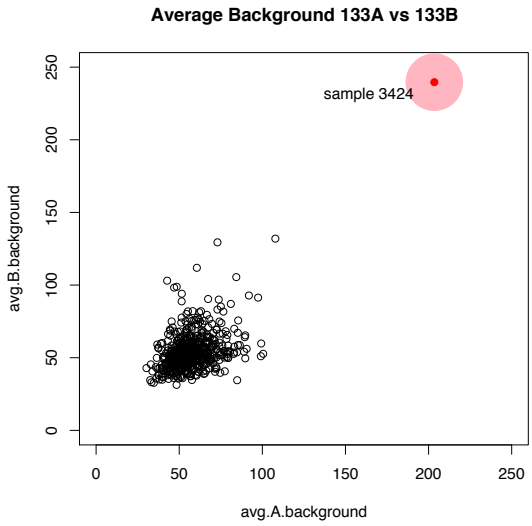


Figure 2.2: Background graph

B.3.3 Percent present

MAS5.0 detection calls (absent, present, marginal) are made for each gene based on the difference between perfect match (PM) and mismatch (MM) probes [m Liu et al., 2002]. While this parameter may be misleading in terms of the absolute value of genes expressed, (as with other parameters) a wildly aberrant value for a particular chip may indicate unintended experimental variation.

Figure 2.3 shows a histogram/distribution of the percent of probesets called 'present' across all chips. Visual inspection of this graph does not suggest outliers. However, Figure 2.4 shows the percent present calls for the A chips plotted against the corresponding values on the B chips. Clearly, Tissue 31754 is dissimilar to the rest of the arrays. Figure 2.4 also demonstrates the utility of comparing the A and B data for outlier detection. While the values for 31754 are not particularly anomalous for either chip singly, the overall 'shape' of the data suggests that Tissue 31754 behaves differently than the rest of the samples. Tissue 31754 was removed from the data set.

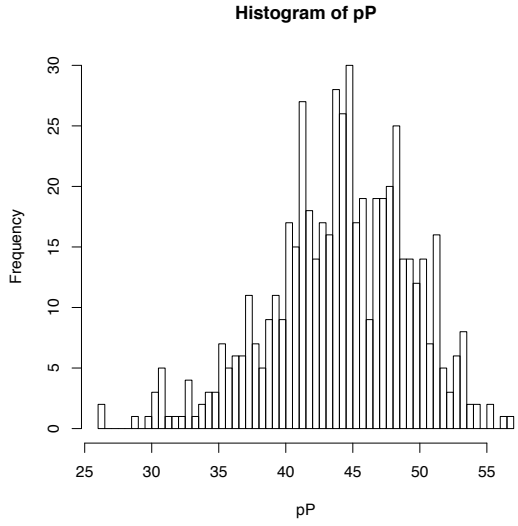


Figure 2.3: Histogram of percent present

B.3.4 Spike-in probesets

According to standard Affymetrix Gene Chip protocols, *e. coli* transcripts *BioB*, *BioC*, *BioD*, and the P1 bacteriophage transcript *CreX* are spiked into the hybridisation solution at increasing concentration to confirm low-end assay sensitivity and appropriate dose response across the dilution range [Affymetrix, 2004a]. Figure 2.5 shows the probeset expression response across all 548 tissues. The observed response clearly does not match the expected linearly increasing

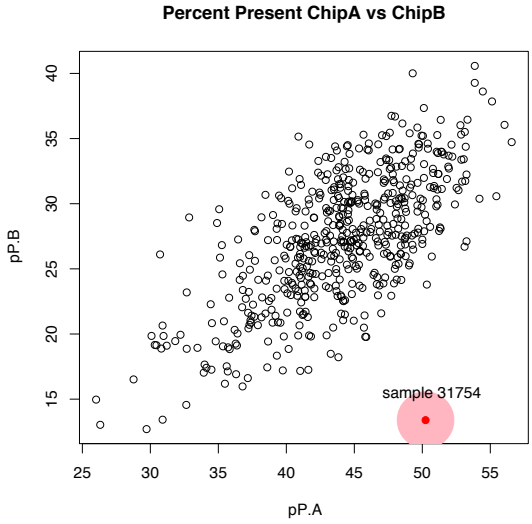


Figure 2.4: Percent present ChipA vs. ChipB

expression values. Correspondence from Gene Logic confirmed that the company does not spike in the bacterial control transcripts as per the Affymetrix guide. No quality assessment could be made from the spike-in controls.

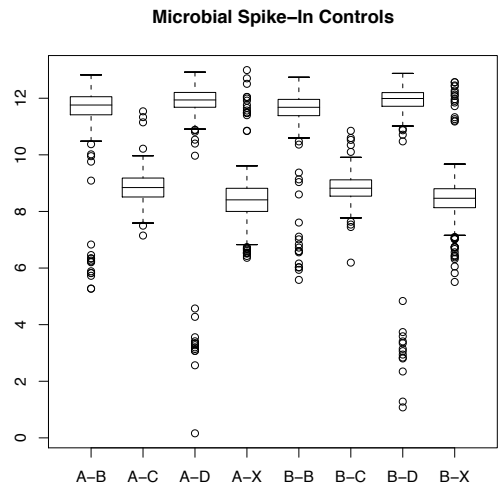


Figure 2.5: Microbial spike-ins

B.3.5 Control probe degradation

Affymetrix probeset sequences are generally chosen to react with approximately the last 600bp (3' terminus) of each gene or EST target transcript [Affymetrix, 2004b]. However, to test transcript efficiency and possible 5'-biased degradation, two 'housekeeping' genes (*GAPDH* and β-actin) are each targeted at three locations along the entire gene transcript. For both of these gene targets, there is one probeset for each of the 3'-transcript tail, mid-transcript, and 5' -transcript head. By comparing the ratio of binding to the 3' tail against the binding to the mid- and 5' transcript, one may gain clues regarding sample transcript quality – at least for these genes.

Figure 2.6 shows the B-actin and *GAPDH* ratios for 3':5' and 3':mid transcripts for chip A vs chip B.

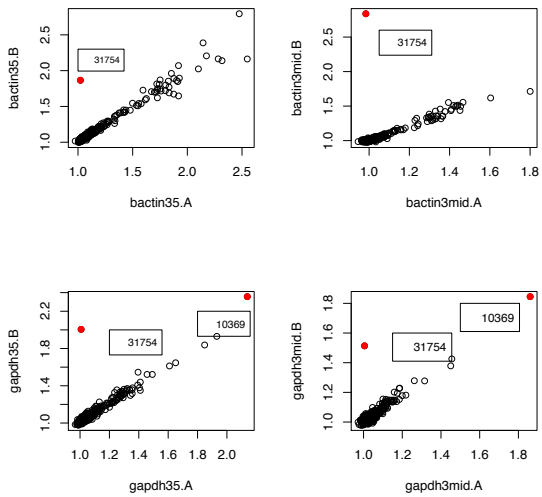


Figure 2.6: QC Probes GAPDH

Inspection of these data suggests that Tissue 31754 has a visibly different ratio profile across both of these genes and the ratios for 10369 are very high for both chip sets. A closer look at the 3'-mid ratio for *GAPDH* shown in Figure 2.7 further reveals that Tissue 10369 should conservatively be designated an outlier.

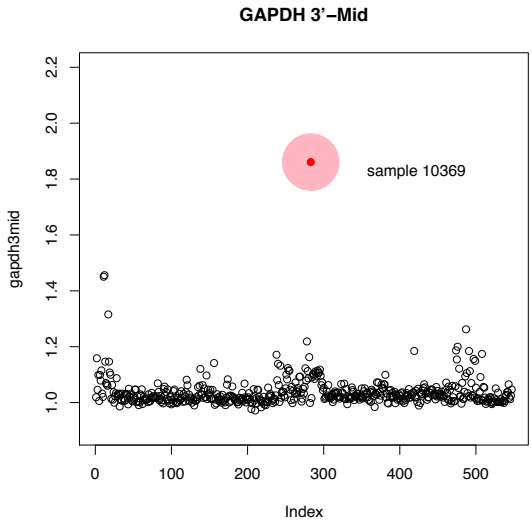


Figure 2.7: Gapdh 3 mid

Tissue 10369 was be removed from the data set and tissue 31754 was previously selected for removal (above).

B.4 RNA degradation analysis

B.4.1 28S:18S ratio

In addition to Affymetrix hybridisation control data, Gene Logic has provided pre-reaction Bioanalyzer analysis of 28S:18S ratios for ribosomal RNA subunit intensities *for some specimens*. The role of ribosomal RNA subunit ratios in the quality control process for microarrays is not clear and the literature is conflicting on their utility. Traditionally, a 28S:18S ratio of 2:1 has been an acceptable ratio for 'good' RNA and this ratio is suggested by Affymetrix [Affymetrix, 2004b]. However, these values are tissue dependent and the ratio has been shown to be dependent on (for example) connective tissue levels, tissue RNase concentration, and whether or not the sample is tumour [Skrypina et al., 2003]. Several studies

suggest that 28S:18S ratios can be misleading and be of "no practical value" [Schoor et al., 2003, Dumur et al., 2004]. Furthermore, at least one study has concluded that these ratios are poorly indicative of the integrity or quality of the RNA sample [Dumur et al., 2004].

Gene Logic internal quality control procedures utilize a significantly lower threshold for this ratio, 0.5 and 1.0 (conflicting correspondence) [GeneLogic, 2005]. Without access to the complete electrophoresis (or Bioanalyzer) chromatogram, 28S:18S values provided by Gene Logic were analyzed and compared to assay-based QC probe (i.e. *GAPDH*, β-actin) results.

Figure 2.8 shows the distribution of 28S:18S results across 400 arrays for which data were provided. Nearly all samples (99%) have ratio values less than the ideal 2:1 ratio and there is considerable variation about the mean (1.245, sd=0.325). There are three samples with ratio values greater than 2.25. While these three samples show discordantly high ratio results, without further information about specific peak profiles there is insufficient evidence for culling such chips from the data set. Finally, the ratio distribution appears truncated at a lower minimum value of 0.5, suggesting that this is the lowest acceptable limit by GeneLogic.

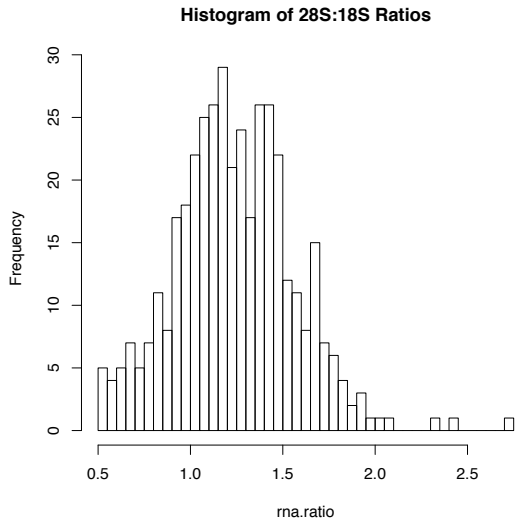


Figure 2.8: Histogram of 28s-18s ratio

B.4.2 Within-probeset degradation

The final technique used to explore potentially problematic tissues was the total-array response for all 11-probe probesets (both PM and MM) across all genes. Generally, each transcript is targeted on the gene chip by 11 discrete (usually non-overlapping) perfect match (and mismatch) 25-mer probes. The mean intensity value for each of these individual probes provides information about the average binding response for all probesets on the chip. Thus, the first probe (#1) reacts with the 5' transcript target while the last probe (#11) reacts with the 3' transcript terminus.

Figure 2.9 shows the mean intensity for all chips at each probe location along the probeset. For illustration, this plot depicts only 20 representative arrays but the expected trend of a high intensity for the 3' probes relative to the 5' probe is readily apparent.

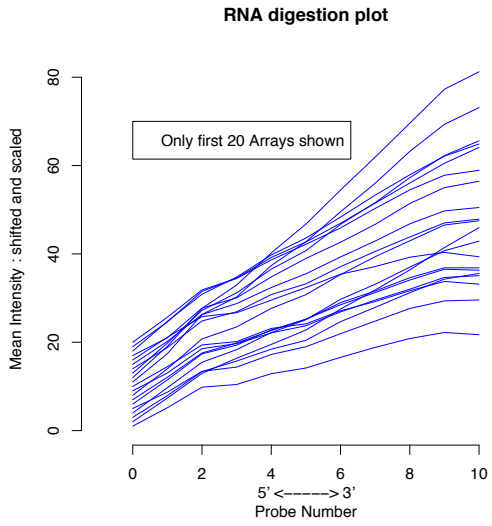


Figure 2.9: RNA degradation plots

Another way to describe this binding trend is to calculate the positive slope for each array observed moving across the probesets (from 1 to 11 or, equivalently from 5' to 3'). Figure 2.10 shows the distribution of slope values across the A

chips; the B chips yield a similar result, data not shown. Interestingly, these data suggest a bimodal distribution with a primary population slope near 2.0 and a secondary population with a higher value between 5.0-6.5. Further investigation identified that most of these high-slope points correspond to a sub-population of arrays hybridised during 2004. This observation is potentially important because the majority of chips (503/548) were hybridised in 2002. See Table B.1.

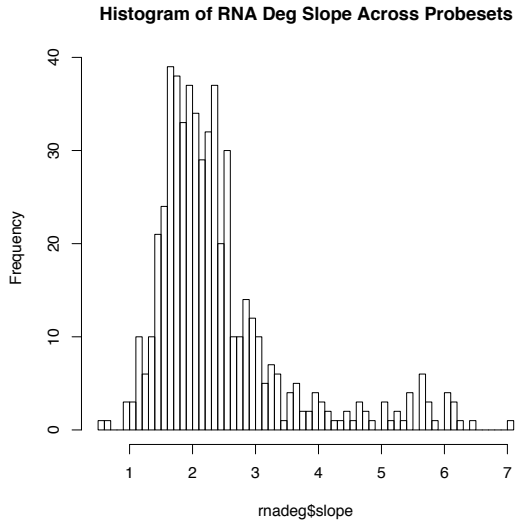


Figure 2.10: RNA degradation plot

Table B.1: Chip hybridisation by year.

	2002	2003	2004
Arrays hybridised	503	17	28

As with the other standard QC metrics analysed above, one can also explore the intra-tissue (or more correctly, the intra-hybridisation solution) response by plotting the A chip slope vs. the B chip slope (Figure 2.11). Again, this technique of viewing intra-tissue response across both chips allows identification of possible outliers. One outlier tissue (31754) was previously identified for removal from the scrubbed data set.

Figure 2.12 shows the same data (degradation slope chip A vs chip B) with highlighting for the 2003 and 2004 chips. As discussed above, we note that the

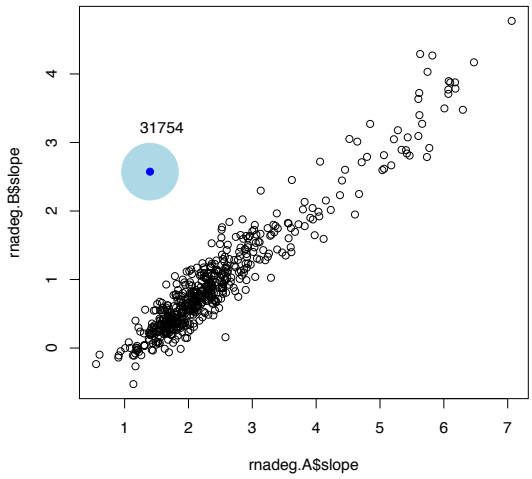


Figure 2.11: RNA slope A vs B

"2004-hybridised" chips are disproportionately represented at the high end of the intra-probeset slopes.

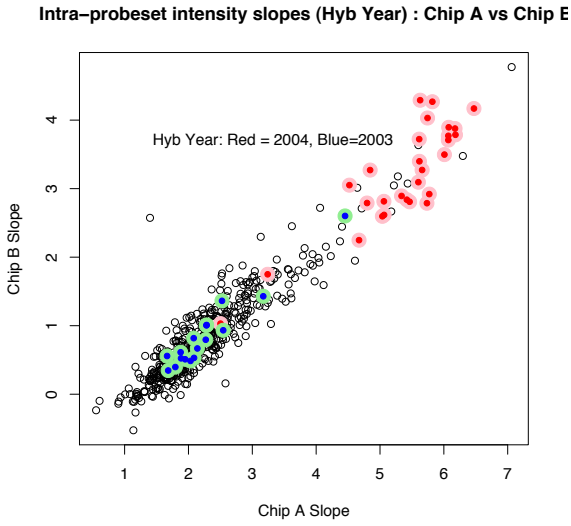


Figure 2.12: Slopes AvB by Hyb Year

Subsequent investigation regarding the twenty-eight "2004" samples identified that these tissues were all processed using a 'Microsample Amplification' proto-

col which is applied to very small amounts of RNA, such as typically recovered in laser capture microarray techniques. In practical terms, this protocol involves a two round of amplification instead of the usual single round. Consequently, these "2004" chips were removed from further analysis.

Finally, the intra-probeset binding slopes provide further perspective on the question raised above regarding the utility of 28S:18S RNA subunit ratios to predict on-chip binding behavior. The *a priori* expectation is that higher intra-probe set binding slopes should be observed for those tissues with relatively poor 28S:18S RNA ratios. Logically, one would expect that tissues with an increased level of RNA degradation will yield lower binding for the 5' transcripts because there is less such product available due to preferential degradation of the 5' transcript. On the other hand, the 3' (with intact poly-A tail) will degrade more slowly and consequently yield a higher probe intensity. One might consequently expect that this (molecular) bias in the degradation process will result in arrays with higher slopes across the 11 25-oligo-mer probes. Figure 2.13 shows the intra-probeset slopes plotted against the 28S:18S ratio for the same tissues.

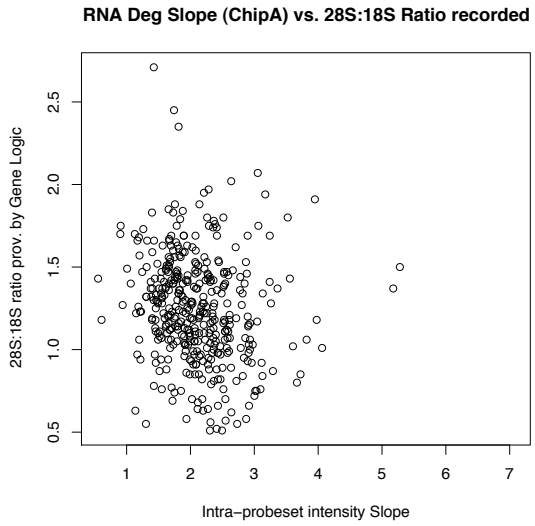


Figure 2.13: RNA deg vs 28s:18s

Visual inspection of Figure 2.13 suggests that there is marginal, if any, corre-

lation between the value of 28S:18S ratio and the resulting intra-probeset ratio moving across the last 600 bases of each transcript. For example the highest slope values (5.0) shown here correspond to a relatively "good" ribosomal subunit ratio (1.5). Further the lowest ribosomal RNA ratios (0.5) do not result in particularly high intra-probeset degradation slopes. These data support the conclusion that ribosomal subunit RNA ratios are poor predictors of on-chip binding behaviour.

B.5 Principal component analysis

Moving beyond the elementary quality analysis involved in outlier array detection, the entire data set was also explored using principal component analysis (PCA). This technique involves attempting to reduce the massively multivariate nature of the data matrix ($N = 548$ samples, $p = 44,928$ probesets) to a new set of uncorrelated (orthogonal) variables that capture the essential variation structure of the data. By visually inspecting the data along the first several principal components, data-wide variance structure may become apparent which can be correlated with experimental conditions. Such structure may be a warning that underlying experimental variation (by design or otherwise) could influence more sophisticated multivariate analysis.

Ultimately, the goal of PCA is to better understand the correlation structure within the data which may then suggest variable relationship hypotheses that can be further investigated [Chatfield and Collins, 1981].

Show in Figure 2.14 is the entire 548 arrays projected onto the first two principal components for the A chips (left) and B chips (right).

Visual inspection of these data hint that there may be two sub-populations of data within the A chips delineated along the second component axis. The B chips plot, on the hand, suggests a single diffuse data cloud in the first to component dimensions. This is interesting bearing in mind that the A chip targets specific or hypothetical gene targets while the B chip contains probesets intended to

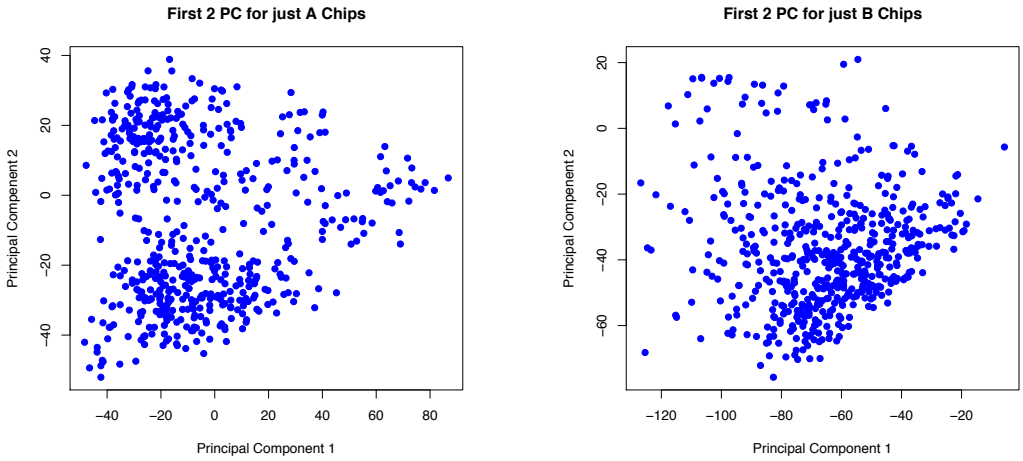


Figure 2.14: PCA all data – just A and just B chips)

hybridise to less well defined expressed sequence tags (ESTs).

B.6 Conclusion

Based on this quality control assessment and data review, four tissues should be conservatively removed from the initial data mining experiments based on outlier analysis applied to the A vs. B chip data. An additional 28 tissues were shown to exhibit high RNA degradation slopes for both the A and B chips. Subsequent communication with the vendor confirmed that these tissues were processed by a "micro-sample amplification" protocol which involves a second round of RNA amplification. Given the possibility of confounding effects of this protocol, these 28 chips were also removed.

Appendix C

Machine learning algorithms

In Chapters 3 and 4 discriminant analysis techniques were introduced based on closed-form analytical solutions to the learning problem of discriminating in the two class-case. The aim of this chapter is to introduce and discuss an iterative, algorithm-based technique called support vector machines (SVMs). Unlike the discriminant techniques introduced in the body of this thesis, SVMs do not attempt to implicitly model the distribution of the data. SVM may have utility in the special case of $p >> n$. In particular, support vector machines has been applied with reasonable success to the field of gene expression analysis [Hastie and Zhu, 2006, Li et al., 2001b].

C.1 Support Vector Machines

To introduce the support vector machine (SVM) algorithm, we first review the genesis of linear learning machines introduced by Rosenblatt and then discuss the modern SVM algorithm. For convenience, we will focus on the two class case as for previous chapters.

The "perceptron" algorithm was developed by Rosenblatt in 1958 to explore models of pattern discrimination, information storage in a biological (and machine) system and recall [Rosenblatt, 1958]. This iterative algorithm is "mistake-driven"

such that for each iteration the linear coefficients $\mathbf{w}$ are updated for each observation $\{\mathbf{x}_i\}$ that is incorrectly classified according to class-separating hyperplane given by Cristianini and Shawe-Taylor [2000],

$$\mathbf{X}\mathbf{w} + b = 0,$$

where $\underset{N\times p}{\mathbf{X}}$ are the data and b is a scalar intercept term. For binary data we code the output targets $y_i \in \{-1, 1\}$ where $i \in \{1, \cdots, N\}$ and run Rosenblatt's perceptron as shown below. After the algorithm converges to a solution $\mathbf{w}$, b we then classify each (future) observation by evaluating the function

$$f(\mathbf{x}_i) = \langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b \tag{C.1}$$

to estimate y_i based on sign(f) as follows:

$$\hat{y}_i = \begin{cases} 1 & \text{iff } f(\mathbf{x}_i) \geq 0, \\ -1 & \text{otherwise.} \end{cases} \tag{C.2}$$

The standard form of the single layer perceptron algorithm is shown below in Algorithm 2 [Cristianini and Shawe-Taylor, 2000].

As presented here the perceptron algorithm will iterate endlessly in the non-separable case as it will not be possible to satisfy the condition $y_i(\langle \mathbf{w} \cdot \mathbf{x} \rangle + b) > 0$ for all i.

We consider this algorithm here as a convenient path to introduce several key elements of more sophisticated techniques such as SVMs. We note, for example, that the product $\gamma_i = y_i f(\mathbf{x}_i)$, known as the *functional margin,* always be positive if $\mathbf{x}_i$ is correctly classified, i.e. the sign of y_i and f agree.

Also, the coefficient $\mathbf{w}$ and intercept b are only updated in the case where the margin γ_i is negative or zero, i.e. $\mathbf{x}_i$ is misclassified. In this case $\mathbf{w}$ is incremented by $\eta y_i \mathbf{x}_i$. Given that coefficients are initially zero (by definition), we can see that $\mathbf{w}$ will necessarily be a linear combination of the observations

$$\mathbf{w} = \sum_{i}^{N} \alpha_i y_i \mathbf{x}_i. \tag{C.3}$$

Algorithm 2 Standard (primal) form of the perceptron algorithm.

For a linearly separable set of observations $\mathbf{X} \in \mathbb{R}^p$ with target outputs $y_i \in \{-1, 1\}$,

Choose learning rate $\eta \in \mathbb{R}^+$.

Initialize: $\mathbf{w} \leftarrow 0;\ b \leftarrow 0;\ R \leftarrow \max \|\mathbf{x}_i\|$, where $i \in \{1, \ldots, N\}$.

Repeat

  for each i

    if $y_i(\langle \mathbf{w} \cdot \mathbf{x} \rangle + b) \leq 0$, then

      $\mathbf{w} \leftarrow \mathbf{w} + \eta y_i \mathbf{x}_i$

      $b \leftarrow b + \eta y_i R^2$

    end if

  end for

until $y_i(\langle \mathbf{w} \cdot \mathbf{x} \rangle + b) > 0$ for all i.

Return $\mathbf{w}, b$.

C.1.1 Wolfe dual

This derivation is particularly useful as it leads to an alternative *dual* representation of the observations, where rather than describing each point $\mathbf{x}_i$ in the original p-dimensional space of the measured data, we can describe the observation in the dual coordinate system of the coefficients. Hand notes that in this dual representation the decision surface becomes a single point while the observations transform from individual points to lines (or hyperplanes) [Hand, 1997]. As the dual space representation are essential aspect of SVMs, it is worth recapitulating the perceptron algorithm given above in the dual form [Cristianini and Shawe-Taylor, 2000], presented in Algorithm 3.

Finally, by substituting the right hand term of Eq.C.3 into the Eq.C.1 we can also rewrite the decision function given in Eq.C.2 as follows

$$\begin{aligned} f(\mathbf{x}_i) &= \langle w \cdot x_i \rangle + b \\ &= \left\langle \sum_j^N \alpha_j y_j \mathbf{x}_j \cdot \mathbf{x}_i \right\rangle + b \\ &= \sum_j^N \alpha_j y_j \langle \mathbf{x}_j \cdot \mathbf{x}_i \rangle + b. \end{aligned}$$

Algorithm 3 Dual form of the Perceptron.

For separable observations $\mathbf{X} \in \mathbb{R}^p$ with target outputs $y_i \in \{-1, 1\}$,

Initialize: $\alpha \leftarrow 0;\ b \leftarrow 0;\ R \leftarrow \max \|\mathbf{x}_i\|$, where $i \in \{1, \cdots, N\}$.

Repeat

for each i

if $y_i(\sum_j^N \alpha_j y_j \langle \mathbf{x}_j \cdot \mathbf{x}_i \rangle + b) \leq 0$, then

$\alpha_i \leftarrow \alpha_i + 1$

$b \leftarrow b + y_i R^2$

end if

end for

until the for loop is correct for all i.

Return α, b.

While both the dual form and primal form of the perceptron are guaranteed to converge in a finite number of iterations for linearly separable classification problems, the resulting solutions are not, however, unique. As with the shortest least squares approach described in the previous chapter, one can force a unique solution by imposing an additional constraint. In the case of the SVM, we choose from the infinite number of separating hyperplanes that solution which has the largest functional margin across all points, i.e. the solution which maximally separates the two classes [Moguerza and Munoz, 2006]. This solution was described by Vapnik as the *optimal hyperplane* and is always unique [Cortes and Vapnik, 1995].

To find the optimal hyperplane we begin by rescaling $\mathbf{x}$ and b such that the functional margin between the decision surface and the subset of observations nearest to this hyperplane from both classes are exactly equal to 1. These observations are hereafter referred to as the *support vectors* [Cortes and Vapnik, 1995]. This scaling recasts the previous solution to yield two parallel hyperplanes called the *canonical hyperplanes* [Burges, 1998, Cristianini and Shawe-Taylor, 2000] such that

$$\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b \geq +1, \quad \text{for } y_i = 1, \text{ and}$$

$$\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b \leq -1, \quad \text{for } y_i = -1.$$

The distance from the positive ($y_i = 1$) (and negative ($y_i = -1$)) support vectors to the decision surface is $1/\left\|\mathbf{w}\right\|_2$, and the distance between the canonical hyperplanes is thus $2/\left\|\mathbf{w}\right\|_2$. Consequently, the maximal margin will be achieved by minimizing $\left\|\mathbf{w}\right\|_2$, the L_2 norm [Cortes and Vapnik, 1995]. Hence we now find the optimal hyperplane solution as given by

$$\begin{aligned} &\text{minimize} \quad \langle \mathbf{w} \cdot \mathbf{w} \rangle, \\ &\text{subject to} \quad y_i \langle w \cdot x_i \rangle + b \geq 1 \text{ for i} \in \{-1, 1\}. \end{aligned}$$

This is a convex optimization with a convex objective function including N simultaneous linear constraints and is solved by introducing N Lagrange multipliers $\alpha_1, \alpha_2, \cdots, \alpha_N$ to construct the (primal) Lagrangian:

$$\begin{aligned} L_P(\mathbf{w}, b, \alpha) &= \frac{1}{2}\langle \mathbf{w} \cdot \mathbf{w} \rangle - \sum_i^N \alpha_i \left[y_i(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b) - 1 \right] \\ &= \frac{1}{2}\langle \mathbf{w} \cdot \mathbf{w} \rangle - \sum_i^N \alpha_i y_i(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b) + \sum_i^N \alpha_i. \end{aligned}$$

We can reformulate this Lagrangian into the Wolfe dual form [Burges, 1998, Platt, 1999] and *maximising* $L(\mathbf{w}, \mathbf{a}, b)$ by differentiating w.r.t. $\mathbf{w}$ and b

$$\begin{aligned} \frac{\partial L(\mathbf{w}, b, \alpha)}{\partial \mathbf{w}} &= \mathbf{w} - \sum_i^N \alpha_i y_i \mathbf{x}_i = 0, \\ \frac{\partial L(\mathbf{w}, b, \alpha)}{\partial b} &= \sum_i^N \alpha_i y_i = 0, \\ \forall \alpha_i &\geq 0 \quad , \; i \in 1, \cdots, N \end{aligned}$$

and solving for the stationary conditions:

$$\mathbf{w} = \sum_i^N \alpha_i y_i \mathbf{x}_i \tag{C.4}$$

$$0 = \sum_i^N \alpha_i y_i \tag{C.5}$$

and re-substituting these relations into Eq.C.4 above to yield

$$L_D(\mathbf{w}, b, \alpha) = \sum_i^N \alpha_i - \frac{1}{2} \sum_{i,j}^N \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i \cdot \mathbf{x}_j \rangle . \tag{C.6}$$

Hence, after solving for α_i (discussed below, see Section C.1.2) we can then calculate $\mathbf{w}$ as in Eq.C.4 which, we note, is unchanged from Eq.C.3.

We note that Eq.C.6 has the interesting property that the data $\mathbf{x}$ only enter the solution through the inner product. This fact will have important implications as we consider the SVM extensions below [Cortes and Vapnik, 1995, Cristianini and Shawe-Taylor, 2000].

C.1.2 Soft margin optimisation

In this simplest implementation of the maximum margin classifier as described so far, the algorithm will not converge if the training data are not linearly separable and the objective function of the Lagrangian dual will increase without bound [Burges, 1998]. To handle the non-separable case we extend the learning machine by introducing a penalty term, also called a *slack variable* [Cortes and Vapnik, 1995] which admits training errors to handle noisy and inseparable data.

We introduce slack variables by augmenting the linear constraints of Eqs.C.4 and C.4,

$$\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b \geq +1 - \xi_i, \quad \text{for } y_i = 1, \text{ and} \tag{C.7}$$

$$\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b \leq -1 + \xi_i, \quad \text{for } y_i = -1 \text{ such that} \tag{C.8}$$

$$\xi_i \geq 0 \quad \forall i, \tag{C.9}$$

and by introducing a cost term to the objective function such as

$$\langle \mathbf{w} \cdot \mathbf{w} \rangle + C \sum_i^N \xi^\sigma,$$

where C is user defined and σ is any positive integer [Burges, 1998]. For a sufficiently large C and sufficiently small σ this regularisation will ensure the hyperplane solution with the minimum mis-classification rate while separating all other (inter-class) observations by the maximum margin [Cortes and Vapnik, 1995]. Vapnik and Cortes introduced the phrase *soft margin optimisation* to describe this "diffusing" effect on the margin by the slack variable [Cortes and

Vapnik, 1995, Cristianini and Shawe-Taylor, 2000]. For computational reasons $\sigma = 1$ (the 1-norm) is often used to avoid the case of NP-completeness and to provide the additional advantage of dropping the slack variable (and their Lagrange multipliers) from the dual form [Vapnik, 1995, Cortes and Vapnik, 1995, Burges, 1998].

The 1-norm soft margin optimisation (also known as the "box constraint" is given by

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2}\langle \mathbf{w} \cdot \mathbf{w} \rangle + C \sum_{i}^{N} \xi_i, \\ &\text{subject to} \quad y_i \left(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b \right) \geq 1 - \xi_i, \; i \in \{1, \ldots, N\}, \\ &\xi_i \geq 0 \quad \forall i. \end{aligned}$$

Introducing Lagrange multipliers we now recast the soft margin optimisation into the primal form as

$$L_P(\mathbf{w}, \alpha, \xi, \mathbf{r}, b) = \frac{1}{2}\langle \mathbf{w} \cdot \mathbf{w} \rangle - \sum_{i}^{N} \alpha_i \left[y_i (\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b) - 1 + \xi_i \right] - \sum_{i}^{N} r_i \xi_i.$$

C.1.3 Importance of regularisation

This soft margin extension has also been shown to have utility even when separability is engineered by mapping the measured data to a higher-dimensional feature-space using more complex kernels (discussed below). By introducing the slack variable we regularize or smooth the resulting decision surface by allowing noisy data points to fall between the canonical hyperplanes and by allowing relatively outlying training points to be misclassified [Cristianini and Shawe-Taylor, 2000].

Hastie and Zhu [2006] argue that regularization, and not the goal of maximum margin discovery underlies SVM success in high dimensional data (such as gene expression microarrays). The optimization problem of Eq.C.1.2 consists of 1) minimizing the loss associated with misclassified observations; and 2) minimising the effects of the roughness penalty on $\mathbf{w}$. Hastie and Zhu note that margin

maximization in a high-dimensional space without regularisation is likely to lead to overfitting and bad generalised performance. Further, one can draw parallels between the regularization of SVMs and the traditional ridge regression method whereby in both cases the directions of smallest variance (of the eigenfunctions in the case of SVMs and eigenvectors in the case of ridge) are shrunk the most [Hastie and Zhu, 2006, Hastie et al., 2001].

We note, however, that the penalty imposed on the 2-norm of $\mathbf{w}$ in ridge in the case of unit ridge penalty ($\lambda = 1$) is equal to to a maximum margin constraint. Hence, in the case of SVMs, enforcing the maximum margin is the source of the regularization and improved generalization. Also, as with shortest least squares method discussed in Chapter 4, enforcing the minimum length of $\mathbf{w}$ again guarantees a unique solution.

Nevertheless, the regularisation objective of SVMs is severely constrained without the cooperative effect achieved by the introduction of the slack variable. By allowing outlier points to be effectively misclassified (without infinite penalty), the slack variables enable constructive penalisation of $\mathbf{w}$ and regularisation leading to reduced solution complexity and better generalisability.

C.1.4 KKT conditions

To be consistent the support vector machine solution requires that the maximum of the dual form Lagrangian w.r.t. α must coincide with the minimum of the linear primal Lagrangian w.r.t. $\mathbf{w}, b$ [Burges, 1998] forming a saddle point in $2N + 1$ hyperspace described by $\{\mathbf{w}, \alpha, b\}$ [Cortes and Vapnik, 1995]. Further, extensions to Lagrange theory introduced by Kuhn and Tucker [1951] provide further convenient checks to ensure that we discover the global optimum solution. In this case the Karush-Kuhn-Tucker (KKT) conditions precisely describe the necessary and sufficient conditions for the optimal solution for optimisation problems such as the SVM whereby the feasible region of the convex objective function is constrained by a set of linear constraints [Burges, 1998]. The KKT conditions are best interpreted in a graphical sense. For any solution $\mathbf{w}, \mathbf{a}, b$,

each observation can only exist in one of two possible locations with respect to the decision surface. When mapped into the feature space, each observation is either

- located in the interior of the convex solution space, or
- located on the decision surface (that is in part defined by the linear constraints) [Cristianini and Shawe-Taylor, 2000].

In the first case, the observation is within the feasible region so the constraints are *inactive* and Fermat's theorem of function minimisation applies, hence $\alpha_i = 0$ [Cristianini and Shawe-Taylor, 2000]. In the second case, the constraint is *active* and thus $\alpha_i \geq 0$ and $y_i(\mathbf{w} \cdot \mathbf{x}_i + b) = 1$. In this case either $\alpha_i = 0$ or $y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1 + \xi_i = 0$.

To ensure an optimal solution for C.15 we can thus ensure that the solution satisfies the KKT conditions, which are here broken down for convenience into:

- the stationarity constraints

$$\frac{\partial L(\mathbf{w}, \alpha, \xi, \mathbf{r}, b)}{\partial \mathbf{w}} = \mathbf{w} - \sum_i^N \alpha_i y_i \mathbf{x}_i = 0, \tag{C.10}$$

$$\frac{\partial L(\mathbf{w}, \alpha, \xi, \mathbf{r}, b)}{\partial \xi} = C - \alpha_i - ri, \tag{C.11}$$

$$\frac{\partial L(\mathbf{w}, \alpha, \xi, \mathbf{r}, b)}{\partial b} = \sum_i^N \alpha_i y_i = 0, \tag{C.12}$$

- the requirement that all Lagrange multipliers are non-negative

$$\begin{aligned} \alpha_i &\geq 0, \\ r_i &\geq 0, \end{aligned}$$

- the necessity for non-negative slack variables

$$\xi_i \geq 0,$$

- and requirement that all solutions fall within the convex boundary, inclusive

$$\alpha_i[y_i(\mathbf{w}\cdot\mathbf{x}_i+b)-1+\xi_i]=0, \tag{C.13}$$

$$r_i\xi_i=0. \tag{C.14}$$

C.1.5 The SVM solution

We estimate the dual form by using the stationarity conditions of C.10, C.11 and C.12 to replace the primal terms of Eq.C.1.2 to yield

$$L_D(\mathbf{w},\alpha,\xi,\mathbf{r},b)=\sum_i^N\alpha_i-\frac{1}{2}\sum_{i,j}^N\alpha_i\alpha_j y_i y_j\left\langle\mathbf{x}_i\cdot\mathbf{x}_j\right\rangle. \tag{C.15}$$

Hence we confirm that the 1-norm solution, $(C\sum_i^N\xi_i^{\sigma=1})$, again returns the dual result of Eq.C.6.

We construct the optimal margin hyperplane, linear in the input space $\mathbf{x}$, by maximizing C.15 w.r.t. α subject to the constraints that

$$0\le\alpha_i\le C$$

$$\textstyle\sum_i^N\alpha_i y_i=0.$$

Finally, the KKT complementarity conditions of Eq.C.13 and C.14 provide a convenient calculation for b for all observations for which $\alpha_i\neq 0$. Burges suggests that taking the mean of the calculated b for all such observations improves numerical stability [Burges, 1998].

C.1.6 Nonlinear learning boundaries

Finally, for completeness, we review the extension of SVMs to nonlinear learning surfaces in variable space. In practice, we find that very high dimensional data such as gene expression microarrays that typically yield between 10^4 and 10^5 measured variable per observation have sufficient features to guarantee perfect

or near-perfect class separability [Hastie and Zhu, 2006]. For instance, we have previously demonstrated that a typical gene expression microarray of 2-class data in 19 observations uncovered over 40,000 unique 2-probeset solutions involving over 20% of the probesets on the chip that perfectly separated the two classes by linear discrimination analysis [LaPointe et al., 2005a].

The extension of SVMs to non-linear decision surfaces with respect to the original variables involves a transformation of the variables in the original p space to the feature space $\wp$:

$$\phi : \mathbb{R}^p \rightarrow \mathbb{R}^\wp.$$

Importantly, the SVM algorithm still aims to build a *linear* classifier [Cortes and Vapnik, 1995] e.g. $f(\mathbf{x})$ in $\wp$ feature space such as

$$f(\mathbf{x}) = \mathbf{w} \cdot \phi(\mathbf{x}) + b,$$

so that Eq.C.4 becomes

$$\mathbf{w} = \sum_i^N y_i \alpha_i \phi(\mathbf{x}_i). \tag{C.16}$$

Substituting C.16 into C.1.6 we see that

$$f(\mathbf{x}) = \sum_i^N y_i \alpha_i \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}) + b$$

and we find that our feature map kernel ϕ (only) contributes through the dot product [Cortes and Vapnik, 1995]. This key observation suggests that there is no need to explicitly evaluate ϕ as we are able to gain the benefit of mapping to a higher dimensional space implicitly through the dot product (e.g. inner product) as in

$$K(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x})^t \phi(\mathbf{y}). \tag{C.17}$$

Sufficient conditions which define K are provided by Mercer's conditions and such "Mercer's kernels" include e.g. [Moguerza and Munoz, 2006]:

- Linear kernels

$$K(\mathbf{x}, \mathbf{y}) = \mathbf{x}^t \mathbf{y},$$

- Polynomial kernels

$$K(\mathbf{x}, \mathbf{y}) = (C + \mathbf{x}^t\mathbf{y})^d,$$

- and Gaussian kernels

$$K(\mathbf{x}, \mathbf{y}) = e^{\frac{-\|\mathbf{x}-\mathbf{y}\|^2}{c}}.$$

The kernels which are suitable for SVMs relate to the set of functions which form a complete vector space in a Hilbert space known as the reproducing kernel Hilbert space (RKHS) [Moguerza and Munoz, 2006]. From an implementation perspective we note that kernels which map into a RKHS are useful because they provide a mapping of the variable space p to a higher dimensional feature space $\wp$ in order to ensure separability (Cover's theorem) [Moguerza and Munoz, 2006]. Additionally, functions in RKHS also provide a "well-behaved" inner-product which allows us to map the kernel space back into the Reals in order to be evaluated in closed form.

C.1.7 Implementation

Finding an SVM solution to minimize the norm of $\mathbf{w}$ subject to linear constraints is a very large quadratic programming (QP) optimization problem when applied to gene expression microarray data. To implement this algorithm herein we utilize Platt's sequential minimum optimization (SMO) methodology. The SMO algorithm first breaks the QP problem into the set smallest possible QP sub-problems and then applies an analytical solution to each "chunk" which significantly improves on traditional numerical methods in terms of computational costs [Platt, 1998, 1999]. The SMO algorithm has also been applied more generally to large quadratic programming tasks such as, for example, the optimization step in penalized logistic regression [Hastie and Zhu, 2004].

C.2 Conclusions

This chapter reviews the support vector machine learning algorithm which has been shown to be useful for analysing gene expression data. Support vector machines enable discovery of the maximum separating hyperplane between classes of interest. A custom implementation of support vector machines was employed in this research which includes an extension to subset selection.

Appendix D

Extended Tables and Figures

D.1 Materials & methods

D.1.1 Covariates provided with GeneLogic data

Table D.1: List of clinical and assay descriptors provided for each HG U133A and B chip purchased from GeneLogic.

genomics_id	sample_type	LCL_notes
sample_site	pathology_morphology	general_sample_description
sample_specific_pathology_type	general_pathologic_category	primary_site
primary_donar_disease	donor_disease_stage	Gender
Age	Race	ratio.28s.18s
beta.actin.medain	gapdh.median	present.calls
percent.present	absent.calls	marginal.cols
Protocol	version	chiptype
chiplot	operator	sampletype
description	project	comments
solutiontype	solutionlot	fluidicsprotocol
a1recoverymix	a1temperature	a1washcycles
mixpera1wash	brecoverymix	btemperature
bwashcycles	mixperbwash	staintemp
firststaintime	a2recoverymix	a2temperature
a2washcycles	mixpera2wash	secondstaintime
thirdstaintime	a3recoverymix	a3temperature
a3washcycles	mixpera3wash	holdingtemp
station	module	hybyear
hybmonth	hybday	pixelsize
filter	scantemp	scanyear
scanmonth	scanday	scannerid
numberofscans	scannertype	Site.Number
Receive.Date		

D.1.2 KEGG gene pathways

Table D.2: List of gene pathways used for GSEA experiments based on the KEGG pathway lists

MAPK sig. path.	Focal adhesion
Regulation of actin cytoskeleton	Calcium sig. path.
Cytokine-cytokine recptn interactn	Neuroactive ligand-recptn interactn
Ubiquitin mediated proteolysis	Tight junction
Wnt sig. path.	Axon guidance
Cell adhesion molecules (CAMs)	Insulin sig. path.
Jak-STAT sig. path.	Purine metab.
Nat. killer cell mediated cytotox.	GnRH sig. path.
Leuk. transendothelial migration	Adherens junction
Prostate cancer	ErbB sig. path.
Gap junction	Glycan structures - biosynth. 1
Cell cycle	Small cell lung cancer
Cell Communication	ECM-recptn interactn
Colorectal cancer	Melanogenesis
Chronic myeloid leukemia	TGF-beta sig. path.
T cell recptn sig. path.	Toll-like recptn sig. path.
Pancreatic cancer	Long-term depression
Phosphatidylinositol sig. system	Ox phosphorylation
Renal cell carcinoma	Glioma
Apoptosis	Fc epsilon RI sig. path.
Long-term potentiation	Melanoma
Hematopoietic cell lineage	VEGF sig. path.
Ag procesng and presntn	Epithelial cell sig. in H. pylori infx
Endometrial cancer	Adipocytokine sig. path.
Non-small cell lung cancer	p53 sig. path.
Pyrimidine metab.	Acute myeloid leukemia
B cell recptn sig. path.	Glycerophospholipid metab.
PPAR sig. path.	Ribosome
Glycan structures - biosynth. 2	Pathogenic E. coli infx - EPEC
Pathogenic E. coli infx - EHEC	mTOR sig. path.
Tyrosine metab.	Starch and sucrose metab.
Tryptophan metab.	Metab. of xenobiotics by cytochrome P450
Hedgehog sig. path.	Notch sig. path.
Inositol phosphate metab.	Type II diabetes mellitus
Complement and coagulation cascades	Bladder cancer
Glycolysis / Gluconeogenesis	Type I diabetes mellitus
Cholera - Infx	Neurodegenerative Diseases
Basal cell carcinoma	Glycerolipid metab.
Androgen and estrogen metab.	Arachidonic acid metab.
Valine, leucine and isoleucine degradtn.	Fatty acid metab.
Dorso-ventral axis formation	ABC transporters - General
Huntington's disease	SNARE interactns in vesic. transport
Butanoate metab.	Fructose and mannose metab.
Taste transduction	N-Glycan biosynth.
Glycine, serine and threonine metab.	Histidine metab.
Sphingolipid metab.	Thyroid cancer
Pyruvate metab.	Olfactory transduction
Folate biosynth.	O-Glycan biosynth.
Lysine degradtn.	Aminoacyl-tRNA biosynth.
Dentatorubropallidoluysian atrophy (DRPLA)	Glutamate metab.
Propanoate metab.	Selenoamino acid metab.
Porphyrin and chlorophyll metab.	Limonene and pinene degradtn.
Linoleic acid metab.	Alzheimer's disease
Bile acid biosynth.	Ether lipid metab.
Glycan structures - degradtn.	Glutathione metab.

Arginine and proline metab.	Phenylalanine metab.
Benzoate degradtn. via CoA ligation	Citrate cycle (TCA cycle)
Basal transcription factors	1- and 2-Methylnaphthalene degradtn.
Alanine and aspartate metab.	Nicotinate and nicotinamide metab.
Galactose metab.	Urea cycle and metab. of amino groups
gamma-Hexachlorocyclohexane degradtn.	Amyotrophic lateral sclerosis (ALS)
Glycosphingolipid biosynth. neo-lactoseries	Pentose phosphate path.
Regulation of autophagy	DNA polymerase
Nitrogen metab.	Glycosylphosphatidylinositol(GPI)-anchor biosynth.
Carbon fixation	Parkinson's disease
Biosynth. of steroids	beta-Alanine metab.
Polyunsaturated fatty acid biosynth.	Aminosugars metab.
Glycosaminoglycan degradtn.	One carbon pool by folate
Proteasome	Naphthalene and anthracene degradtn.
Alkaloid biosynth. II	RNA polymerase
Heparan sulfate biosynth.	Maturity onset diabetes of the young
Aminophosphonate metab.	Chondroitin sulfate biosynth.
Keratan sulfate biosynth.	Pantothenate and CoA biosynth.
Glycosphingolipid biosynth. - ganglioseries	Riboflavin metab.
Prion disease	Circadian rhythm
Pentose and glucuronate interconversions	Glycosphingolipid biosynth. globoseries
Methionine metab.	Caprolactam degradtn.
Glycosphingolipid biosynth. - lactoseries	Bisphenol A degradtn.
Renin-angiotensin system	3-Chloroacrylic acid degradtn.
alpha-Linolenic acid metab.	Sulfur metab.
N-Glycan degradtn.	Cysteine metab.
Taurine and hypotaurine metab.	Reductve carboxylate cycle (CO2 fixation)
Valine, leucine and isoleucine biosynth.	Methane metab.
Glyoxylate and dicarboxylate metab.	Cyanoamino acid metab.
Terpenoid biosynth.	Phenylalanine, tyrosine and tryptophan biosynth.
Fatty acid elongation in mitochondria	Protein export
Synthesis and degradtn. of ketone bodies	Thiamine metab.
C21-Steroid hormone metab.	D-Glutamine and D-glutamate metab.
Fatty acid biosynth.	Ascorbate and aldarate metab.
Phenylpropanoid biosynth.	Streptomycin biosynth.
Tetrachloroethene degradtn.	Alkaloid biosynth. I
Nucleotide sugars metab.	Biotin metab.
Vitamin B6 metab.	Ubiquinone biosynth.
Caffeine metab.	Geraniol degradtn.
Atrazine degradtn.	Lysine biosynth.
Monoterpenoid biosynth.	Styrene degradtn.
Retinol metab.	Inositol metab.
Novobiocin biosynth.	Peptidoglycan biosynth.
1,4-Dichlorobenzene degradtn.	Fluorobenzoate degradtn.
C5-Branched dibasic acid metab.	Lipoic acid metab.
D-Arginine and D-ornithine metab.	

D.1.3 Gene sets used for GSEA analysis

Table D.3: List of gene symbols used in manually curated Wnt target list. The list is built combining human gene symbols curated by R. Nusse (Stanford Univ, USA) with additional genes identified in the literature review described herein.

CommonName	Symbol	Source
MDR1	ABCB1	RNUSSE
HATH1	ATOH1	RNUSSE
Axin-2	AXIN2	RNUSSE
osteocalcin	BGLAP	RNUSSE
survivin	BIRC5	RNUSSE
BMP4	BMP4	RNUSSE
betaTrCP	BTRC	RNUSSE
MCP-3	CCL7	LCL
cyclin D	CCND1	RNUSSE
CD44	CD44	RNUSSE
E-cadherin	CDH1	RNUSSE
P16ink4A	CDKN2A	RNUSSE
CDX1	CDX1	RNUSSE
CDX4	CDX4	RNUSSE
Claudin-1	CLDN1	RNUSSE
CCN1	CYR61	RNUSSE
Dickkopf	DKK1	RNUSSE
Delta-like 1	DLL1	RNUSSE
EDA	EDA	RNUSSE
endothlin-1	EDN1	RNUSSE
EPHB	EFNB1	RNUSSE
EGF receptor	EGFR	RNUSSE
ENC1	ENC1	LCL
autotaxin	ENPP2	RNUSSE
NBL4	EPB41L4A	LCL
FGF18	FGF18	RNUSSE
FGF20	FGF20	RNUSSE
FGF4	FGF4	RNUSSE
FGF9	FGF9	RNUSSE
fra-1	FOSL1	RNUSSE
FOXN1	FOXN1	RNUSSE
follistatin	FST	RNUSSE
frizzled 7	FZD7	RNUSSE
Gastrin	GAST	RNUSSE
Proglucagon	GCG	RNUSSE
Gremlin	GREM1	RNUSSE
Tcf-1	HNF1A	RNUSSE
ID2	ID2	RNUSSE
IGF-1	IGF1	RNUSSE
IGF-II	IGF2	RNUSSE

IL6	IL6	RNUSSE
IL8	IL8	RNUSSE
IRX3	IRX3	RNUSSE
jagged1	JAG1	RNUSSE
c-jun	JUN	RNUSSE
L1 neural adhesion	L1CAM	RNUSSE
LEF1	LEF1	RNUSSE
LGR5/GPR49	LGR5	RNUSSE
MET	MET	RNUSSE
MMP2	MMP2	RNUSSE
MMP26	MMP26	RNUSSE
stromelysin	MMP3	RNUSSE
MMP-7	MMP7	RNUSSE
MMP9	MMP9	RNUSSE
c-myc	MYC	RNUSSE
c-myc binding protein	MYCBP	RNUSSE
nanog	NANOG	RNUSSE
neurogenin1	NEUROG1	RNUSSE
Nkx2.2	NKX2-2	RNUSSE
NOS2	NOS2A	RNUSSE
NrCAM	NRCAM	RNUSSE
uPAR	PLAUR	RNUSSE
perostin	POSTn	RNUSSE
PPARdelta	PPARD	RNUSSE
cyclooxygenase	PTGS2	RNUSSE
PTTG	PTTG1	RNUSSE
RET	RET	RNUSSE
Wrch1	RHOU	RNUSSE
RUNX2	RUNX2	RNUSSE
SALL4	SALL4	RNUSSE
SIX3	SIX3	RNUSSE
SOX2	SOX2	RNUSSE
SOX9	SOX9	RNUSSE
Brachyury	TBX1	RNUSSE
ITF-2	TCF7L2	RNUSSE
TIAM1	TIAM1	RNUSSE
ZO-1	TJP1	LCL
RANK ligand	TNFSF11	RNUSSE
Twist	TWIST1	RNUSSE
versican	VCAN	RNUSSE
VEGF-C	VEGC	RNUSSE
VEGF	VEGF	RNUSSE
WISP	WISP	RNUSSE
WISP-1	WISP1	RNUSSE
WISP-2	WISP2	RNUSSE
sFRP-2	WNT4	RNUSSE

D.2 Normal tissue analysis

D.2.1 Genes elevated in proximal tissues

				Proximal-Distal			Cecum-Rectum			Validation			
Rank	Probeset ID	Symbol	Description	Expr. Δ	t	P-Value	Expr. Δ	t	P-Value	P-Value	t	CI Low	CI High
1	222262_s_at	ETNK1	ethanolamine kinase 1	3.3492	-12.9258	5.27E-23	3.5741	-9.0521	6.53E-09	1.37E-01	1.5891	-0.3764	2.4320
2	225458_at	SEC6L1	SEC6-like 1 (S. cerevisiae)	5.4422	-12.5937	5.10E-22	6.2917	-9.2685	2.57E-09	1.75E-01	1.4370	-0.7340	3.6253
3	225457_s_at	SEC6L1	SEC6-like 1 (S. cerevisiae)	4.2221	-12.5347	7.62E-22	4.9764	-9.7261	3.59E-10	2.19E-01	1.2930	-0.8902	3.5413
4	219017_at	ETNK1	ethanolamine kinase 1	4.0801	-12.3947	1.98E-21	4.1238	-8.1023	3.99E-07	2.63E-01	1.1704	-1.0423	3.4942
5	207558_s_at	PITX2	paired-like homeodomain transcription factor 2	1.6252	-12.3516	2.66E-21	1.7549	-8.5481	5.79E-08	5.20E-01	0.6582	-0.6362	1.2099
6	224453_s_at	ETNK1	ethanolamine kinase 1	2.0637	-11.5429	6.45E-19	2.1692	-8.0763	4.47E-07	2.07E-01	1.3638	-0.1907	0.7586
7	229230_at	OSTalpha	organic solute transporter alpha	2.4793	-10.8011	9.47E-17	2.7768	-8.6246	4.15E-08	1.95E-01	1.3510	-0.4902	2.2212
8	206340_at	NR1H4	nuclear receptor subfamily 1, group H, member 4	2.0505	-10.3266	2.22E-15	2.4066	-9.1541	4.20E-09	3.55E-02	2.3580	0.0394	0.9527
9	226432_at		** no description **	2.3181	-10.0408	1.46E-14	2.5744	-7.2261	1.76E-05	2.49E-01	1.2193	-0.5313	1.8442
10	209869_at	ADRA2A	adrenergic, alpha-2A-, receptor	1.6585	-9.8367	5.55E-14	1.7705	-8.0507	4.99E-07	2.45E-01	1.2272	-0.4738	1.6677
11	227194_at	FAM3B	family with sequence similarity 3, member B	2.8282	-9.8079	6.70E-14	3.4326	-6.9816	5.00E-05	2.04E-01	1.3699	-0.6662	2.7145
12	207251_at	MEP1B	meprin A, beta	1.7581	-9.7239	1.16E-13	1.8022	-6.5673	2.91E-04	1.52E-01	1.5371	-0.2025	1.1482
13	219954_s_at	GBA3	glucosidase, beta, acid 3 (cytosolic)	1.7033	-9.6737	1.60E-13	1.9800	-8.3619	1.30E-07	1.76E-01	1.4742	-0.2567	1.1929
14	219955_at	FLJ10884	hypothetical protein FLJ10884	1.8400	-9.1831	3.77E-12	1.9031	-5.9016	4.66E-03	2.78E-01	1.1257	-0.0917	0.2976
15	225290_at		** no description **	2.2680	-9.1191	5.68E-12	2.4516	-6.2630	1.04E-03	3.30E-01	1.0125	-0.8929	2.4715
16	201920_at	SLC20A1	solute carrier family 20 (phosphate transporter), member 1	2.1030	-8.5555	1.97E-10	2.3428	-7.0466	3.79E-05	3.68E-01	0.9338	-1.0459	2.6359
17	206294_at	HSD3B2	hydroxy-delta-5-steroid dehydrogenase, 3 beta- and steroid delta-isomerase 2	1.8455	-8.2334	1.43E-09	2.0613	-6.6283	2.25E-04	3.68E-01	0.9331	-0.9742	2.4564
18	231576_at		** no description **	2.1646	-8.0045	5.75E-09				1.89E-01	1.4363	-0.3026	1.3050
19	222943_at	GBA3	glucosidase, beta, acid 3 (cytosolic)	2.0596	-7.9083	1.03E-08	2.5806	-6.9404	5.96E-05	3.62E-01	0.9560	-0.7354	1.8413
20	202236_s_at	SLC16A1	solute carrier family 16 (monocarboxylic acid transporters), member 1	1.6747	-7.6989	3.58E-08	1.8552	-6.9860	4.91E-05	7.30E-01	-0.3520	-1.4137	1.0142
21	205366_s_at	HOXB6	homeo box B6	1.4861	-7.6727	4.18E-08	1.6332	-6.0387	2.65E-03	3.75E-01	0.9368	-0.3720	0.8890
22	222774_s_at	NETO2	neuropilin (NRP) and tolloid (TLL)-like 2	1.6919	-7.5826	7.11E-08				6.56E-01	0.4551	-0.5353	0.8246
23	235733_at		** no description **	1.1776	-7.4926	1.21E-07	1.2384	-6.0872	2.17E-03	7.99E-02	1.8733	-0.0196	0.3111
24	202235_at	AFARP1	AKR7 family pseudogene	1.2859	-7.3793	2.33E-07	1.3698	-6.6895	1.73E-04	5.44E-01	-0.6204	-0.9183	0.5044
25	224476_s_at	MESP1	mesoderm posterior 1	1.2840	-7.2589	4.68E-07				2.16E-01	1.2876	-0.0855	0.3497
26	206858_s_at	HOXC6	homeo box C6	1.2640	-7.1875	7.05E-07	1.3672	-6.2775	9.82E-04	1.49E-01	1.5380	-0.1110	0.6535
27	208126_s_at	CYP2C18	cytochrome P450, family 2, subfamily C, polypeptide 18	1.5721	-7.0842	1.27E-06				7.70E-01	0.2970	-0.8071	1.0692
28	207529_at	DEFA5	defensin, alpha 5, Paneth cell-specific	2.8342	-7.0313	1.71E-06	3.8363	-5.9701	3.51E-03	1.76E-01	1.5002	-0.4189	1.8957
29	209692_at	EYA2	eyes absent homolog 2 (Drosophila)	1.3808	-6.9744	2.36E-06	1.4435	-5.9334	4.09E-03	2.40E-02	2.5104	0.0383	0.4702
30	214595_at	KCNG1	potassium voltage-gated channel, subfamily G, member 1	1.1633	-6.9706	2.41E-06	1.2868	-6.4306	5.17E-04	9.41E-02	-1.7744	-0.5220	0.0453
31	202888_s_at	ANPEP	alanyl (membrane) aminopeptidase (aminopeptidase N, aminopeptidase M, microsomal aminopeptidase, CD13, p150)	2.6011	-6.8676	4.30E-06	3.3179	-5.7250	9.58E-03	2.63E-01	1.1662	-0.9121	3.0790
32	202718_at	IGFBP2	insulin-like growth factor binding protein 2, 36kDa	1.8892	-6.8559	4.59E-06				7.97E-01	0.2631	-1.0565	1.3500
33	221804_s_at	FAM45A	family with sequence similarity 45, member A	1.3071	-6.8456	4.86E-06				6.85E-01	-0.4156	-1.7005	1.1551
34	207158_at	APOBEC1	apolipoprotein B mRNA editing enzyme, catalytic polypeptide 1	1.4298	-6.7384	8.81E-06				8.55E-01	0.1857	-0.5250	0.6260
35	230949_at	SLC23A3	solute carrier family 23 (nucleobase transporters), member 3	1.1622	-6.5961	1.92E-05				6.05E-02	2.0879	-0.0267	1.0424
36	205541_s_at	GSPT2	G1 to S phase transition 2	1.3378	-6.5339	2.70E-05	1.4485	-5.7155	9.96E-03	1.91E-01	1.4047	-0.2567	1.1282
37	207212_at	SLC9A3	solute carrier family 9 (sodium/hydrogen exchanger), isoform 3	1.2571	-6.5310	2.74E-05				9.52E-01	0.0608	-0.2994	0.3171
38	215103_at	CYP2C18	cytochrome P450, family 2, subfamily C, polypeptide 18	1.3638	-6.5193	2.92E-05	1.4312	-5.9261	4.21E-03	9.81E-01	0.0248	-0.6717	0.6874
39	206755_at	CYP2B6	cytochrome P450, family 2, subfamily B, polypeptide 6	1.2980	-6.4787	3.64E-05	1.3244	-5.5367	2.05E-02	7.86E-03	3.3120	0.1017	0.5198
40	239656_at		** no description **	1.1506	-6.4761	3.69E-05				5.91E-01	0.5545	-0.3367	0.5611
41	222955_s_at	FAM45A	family with sequence similarity 45, member A	1.2688	-6.4573	4.09E-05				8.98E-01	0.1300	-0.2480	0.2802
42	213181_s_at	MOCS1	molybdenum cofactor synthesis 1	1.1617	-6.4528	4.19E-05	1.2410	-6.4040	5.78E-04	8.98E-01	0.1300	-0.2891	0.3268
43	205522_at	HOXD4	homeo box D4	1.2966	-6.4496	4.26E-05	1.4206	-5.6334	1.39E-02	1.70E-02	2.7802	0.0674	0.5621
44	221304_at	UGT1A8	UDP glycosyltransferase 1 family, polypeptide A8	1.3599	-6.4054	5.40E-05				3.32E-02	2.4124	0.0157	0.3156
45	205660_at	OASL	2'-5'-oligoadenylate synthetase-like	1.5483	-6.3676	6.61E-05				9.13E-02	1.8836	-0.1619	1.8170
46	218888_s_at		** no description **	1.6234	-6.3647	6.71E-05				8.65E-01	0.1729	-0.7162	0.8440
47	209900_s_at	SLC16A1	solute carrier family 16 (monocarboxylic acid transporters), member 1	1.4721	-6.3225	8.41E-05	1.6899	-6.0457	2.57E-03	7.73E-01	-0.2938	-1.3553	1.0276
48	242059_at		** no description **	1.6676	-6.3073	9.12E-05				1.58E-01	1.5283	-0.3359	1.7837
49	221305_s_at	UGT1A8	UDP glycosyltransferase 1 family, polypeptide A8	1.6300	-6.3057	9.20E-05				1.16E-01	1.7472	-0.0934	0.7101
50	219197_s_at	SCUBE2	signal peptide, CUB domain, EGF-like 2	1.2723	-6.2538	1.21E-04	1.5426	-7.2700	1.45E-05	1.51E-01	1.5707	-0.0850	0.4708
51	236860_at	NPY6R	neuropeptide Y receptor Y6 (pseudogene)	1.1988	-6.2070	1.55E-04				1.50E-01	1.5108	-0.0514	0.3088
52	218739_at	ABHD5	abhydrolase domain containing 5	1.2190	-6.2061	1.56E-04				8.25E-01	0.2256	-0.4494	0.5557
53	210797_s_at	OASL	2'-5'-oligoadenylate synthetase-like	1.4082	-6.1890	1.70E-04				2.62E-01	1.1791	-0.1607	0.5374
54	206754_s_at	CYP2B6	cytochrome P450, family 2, subfamily B, polypeptide 6	1.5418	-6.1369	2.24E-04				2.00E-01	1.3404	-0.3312	1.4532
55	203333_at	KIFAP3	kinesin-associated protein 3	1.2568	-6.1317	2.30E-04				5.92E-01	0.5550	-0.6324	1.0488
56	224454_at	ETNK1	ethanolamine kinase 1	1.1406	-6.1181	2.47E-04				3.33E-01	0.9980	-0.1088	0.3037
57	214651_s_at	HOXA9	homeo box A9	1.4981	-6.0474	3.57E-04	1.6730	-5.8388	6.02E-03	7.54E-01	0.3192	-0.9026	1.2175
58	242683_at	na	hypothetical gene supported by AK095347	1.2426	-5.9201	6.86E-04				3.97E-02	2.3200	0.0201	0.6997
59	236894_at		** no description **	1.3679	-5.8885	8.07E-04				6.22E-01	0.5028	-0.1866	0.3029
60	218136_s_at	MSCP	mitochondrial solute carrier protein	1.2016	-5.8872	8.12E-04				3.93E-01	0.8820	-0.1419	0.3403
61	210153_s_at	ME2	malic enzyme 2, NAD(+)-dependent, mitochondrial	1.2047	-5.8498	9.82E-04				6.28E-01	0.5001	-0.4716	0.7442
62	209752_at	REG1A	regenerating islet-derived 1 alpha (pancreatic stone protein, pancreatic thread protein)	2.7216	-5.8414	1.02E-03				5.62E-01	-0.5914	-0.3380	0.1901
63	238638_at	SLC37A2	solute carrier family 37 (glycerol-3-phosphate transporter), member 2	1.3919	-5.8351	1.06E-03				5.80E-01	0.5732	-0.5148	0.8685
64	214421_x_at	CYP2C9	cytochrome P450, family 2, subfamily C, polypeptide 9			6.79E-03	1.3877	-5.8095	6.79E-03	8.26E-02	1.8529	-0.0292	0.4316
65	205815_at	PAP	pancreatitis-associated protein	2.0272	-5.7979	1.28E-03	2.7965	-5.5114	2.27E-02	1.36E-01	1.6661	-0.1684	1.0163
66	225351_at	FAM45A	family with sequence similarity 45, member A	1.2592	-5.6944	2.14E-03				8.22E-01	-0.2296	-0.9944	0.8026
67	243669_at	PRAP1	proline-rich acidic protein 1	1.4986	-5.6740	2.37E-03				4.66E-01	0.7466	-0.7334	1.5338
68	228564_at	LOC375295	hypothetical gene supported by BC013438	1.1976	-5.6664	2.47E-03				5.38E-02	2.1149	-0.0035	0.3785
69	223541_at	HAS3	hyaluronan synthase 3	1.4178	-5.6557	2.60E-03				3.82E-01	-0.8990	-1.3977	0.5637
70	202234_s_at	AFARP1	AKR7 family pseudogene	1.4304	-5.6464	2.72E-03				7.49E-01	0.3259	-1.0571	1.4355
71	203920_at	NR1H3	nuclear receptor subfamily 1, group H, member 3			1.34E-02	1.3409	-5.5600	1.87E-02	4.58E-01	0.7617	-0.3137	0.6637
72	231897_at	ZNF483	zinc finger protein 483	1.3192	-5.5272	4.90E-03				9.53E-01	0.0602	-1.1123	1.1762
73	228155_at	C10orf58	chromosome 10 open reading frame 58	1.4264	-5.5143	5.21E-03				8.53E-01	0.1888	-1.3883	1.6572
74	206601_s_at	HOXD3	homeo box D3	1.1325	-5.5056	5.44E-03	1.2135	-5.5679	1.81E-02	3.90E-01	0.8826	-0.1434	0.3488
75	215913_s_at	GULP1	GULP, engulfment adaptor PTB domain containing 1			2.39E-02	1.4578	-5.4985	2.39E-02	2.46E-02	2.4689	0.0299	0.3831
76	208596_s_at	UGT1A3	UDP glycosyltransferase 1 family, polypeptide A3	1.6580	-5.3741	1.03E-02				3.94E-01	0.8810	-0.5799	1.3851
77	202495_at	TBCC	tubulin-specific chaperone c	1.1465	-5.3411	1.20E-02				8.85E-01	0.1471	-0.3784	0.4337
78	221920_s_at	MSCP	mitochondrial solute carrier protein	1.1893	-5.3370	1.23E-02				3.19E-01	1.0688	-0.2442	0.6546
79	223058_at	C10orf45	chromosome 10 open reading frame 45	1.3829	-5.3188	1.34E-02				9.93E-01	0.0092	-1.2206	1.2307
80	219926_at	POPDC3	popeye domain containing 3	1.1296	-5.2863	1.56E-02				1.73E-01	1.4622	-0.0737	0.3604
81	210154_at	ME2	malic enzyme 2, NAD(+)-dependent, mitochondrial	1.3016	-5.2581	1.78E-02				4.06E-01	0.8804	-0.4040	0.8951
82	220753_s_at	CRYL1	crystallin, lambda 1	1.2752	-5.2392	1.95E-02				9.42E-01	0.0735	-0.9931	1.0643
83	205505_at	GCNT1	glucosaminyl (N-acetyl) transferase 1, core 2 (beta-1,6-N-acetylglucosaminyltransferase)	1.1227	-5.2361	1.98E-02				1.91E-01	1.3736	-0.0833	0.3805
84	219640_at	CLDN15	claudin 15	1.1692	-5.2276	2.06E-02				3.03E-01	1.0642	-0.1625	0.4894
85	214038_at	CCL8	chemokine (C-C motif) ligand 8	1.6140	-5.2067	2.27E-02				1.29E-01	1.7169	-0.2431	1.5559
86	220017_x_at	CYP2C9	cytochrome P450, family 2, subfamily C, polypeptide 9	1.3983	-5.1902	2.46E-02	1.5251	-5.4185	3.29E-02	1.56E-03	3.8998	0.1592	0.5472
87	206407_s_at	CCL13	chemokine (C-C motif) ligand 13	1.4448	-5.1730	2.66E-02				9.06E-02	1.8234	-0.0265	0.3189
88	220585_at	FLJ22761	hypothetical protein FLJ22761	1.1558	-5.1501	2.96E-02				7.05E-01	0.3868	-0.1662	0.2388
89	217085_at	SLC14A2	solute carrier family 14 (urea transporter), member 2	1.2940	-5.1161	3.47E-02				1.69E-01	1.5324	-0.3248	1.5282
90	205208_at	FTHFD	formyltetrahydrofolate dehydrogenase	1.2531	-5.1123	3.53E-02				7.99E-01	0.2585	-0.3126	0.3997
91	203639_s_at	FGFR2	fibroblast growth factor receptor 2 (bacteria-expressed kinase, keratinocyte growth factor receptor, craniofacial dysostosis 1, Crouzon syndrome, Pfeiffer syndrome, Jackson-Weiss syndrome)	1.2760	-5.0917	3.89E-02				3.02E-01	1.0705	-0.1918	0.5747
92	204663_at	ME3	malic enzyme 3, NADP(+)-dependent, mitochondrial	1.1447	-5.0447	4.83E-02				5.46E-01	0.6203	-0.3844	0.6922
93	211776_s_at	EPB41L3	erythrocyte membrane protein band 4.1-like 3	1.2553	-5.0391	4.95E-02				5.81E-01	0.5706	-0.4283	0.7236

Figure 4.1: Genes and probesets elevated in proximal tissues relative to distal tissues.

D.2.2 Genes elevated in distal tissues

				Proximal-Distal			Cecum-Rectum			Validation			
Rank	Probeset ID	Symbol	Description	Expr. Δ	t	P-Value	Expr. Δ	t	P-Value	P-Value	t	CI Low	CI High
1	230784_at	PRAC	small nuclear protein PRAC	10.3887	16.6750	4.56E-34	15.5666	18.2177	2.94E-24	1.22E-03	-3.8956	-3.4130	-1.0114
2	230105_at		** no description **	2.2919	12.3536	2.62E-21	2.9669	11.1548	8.54E-13	3.09E-03	-3.6184	-2.1466	-0.5423
3	209844_at	HOXB13	homeo box B13	2.4103	12.1639	9.54E-21	3.1342	10.6863	6.07E-12	6.44E-02	-1.9822	-1.0329	0.0336
4	222571_at	SIAT7F	sialytransferase 7 ((alpha-N-acetylneuraminyl 2,3-betagalactosyl-1,3)-N-acetyl galactosaminide alpha-2,6-sialytransferase) F	1.7332	12.0297	2.38E-20	1.9083	9.5206	8.68E-10	1.74E-02	-2.6361	-1.5450	-0.1712
5	203892_at	WFDC2	WAP four-disulfide core domain 2	2.0622	11.7522	1.56E-19	2.3090	9.5105	9.06E-10	7.58E-02	-1.9010	-0.9904	0.0547
6	214598_at	CLDN8	claudin 8	4.4296	10.9279	4.05E-17	5.9352	9.2485	2.80E-09	2.97E-05	-5.8917	-3.8620	-1.8099
7	230360_at	COLM	collomin	2.1190	10.9209	4.25E-17	2.7368	10.0265	9.94E-11	8.76E-03	-3.1862	-2.8211	-0.5144
8	221091_at	INSL5	insulin-like 5	3.3289	10.2037	5.00E-15	5.0245	9.2341	2.98E-09	2.96E-01	-1.0788	-1.7982	0.5831
9	221164_x_at	CHST5	carbohydrate (N-acetylglucosamine 6-O) sulfotransferase 5	1.5826	9.8032	6.90E-14	1.7349	8.1540	3.19E-07	7.03E-02	-1.9631	-1.2320	0.0559
10	229254_at	DKFZp761N11	hypothetical protein DKFZp761N1114	2.3718	9.5776	2.99E-13	3.0443	9.2865	2.38E-09	1.74E-02	-2.6380	-2.2971	-0.2546
11	230269_at		** no description **	1.8860	9.5192	4.36E-13	2.1495	7.9354	8.23E-07	1.84E-03	-3.7893	-3.0771	-0.8590
12	223942_x_at	CHST5	carbohydrate (N-acetylglucosamine 6-O) sulfotransferase 5	1.5910	9.3437	1.35E-12	1.7763	8.2351	2.25E-07	1.56E-02	-2.7784	-1.2593	-0.1582
13	230845_at	PRAC2	prostate/rectrum and colon protein no. 2	1.2645	9.1328	5.20E-12	1.2799	6.5300	3.40E-04	7.34E-01	-0.3473	-0.4016	0.2897
14	239994_at		** no description **	1.7691	8.9650	1.51E-11	2.1086	7.9228	8.69E-07	3.77E-02	-2.3472	-0.9050	-0.0315
15	40284_at	FOXA2	forkhead box A2	1.3520	8.5397	2.17E-10	1.4577	7.3722	9.37E-06	2.71E-01	-1.1395	-0.6620	0.1987
16	207249_s_at	SLC28A2	solute carrier family 28 (sodium-coupled nucleoside transporter), member 2	2.0334	8.5384	2.19E-10	2.6495	6.8463	8.90E-05	2.60E-01	-1.1847	-0.9239	0.2760
17	242372_s_at	DKFZp761N11	hypothetical protein DKFZp761N1114	1.5715	8.4149	4.70E-10	1.8751	7.5943	3.60E-06	5.96E-02	-2.0524	-0.4335	0.0098
18	213994_s_at	SPON1	spondin 1, extracellular matrix protein	1.6341	8.3820	5.75E-10	1.8277	7.5849	3.75E-06	8.11E-02	-1.8548	-1.3333	0.0858
19	205185_at	SPINK5	serine protease inhibitor, Kazal type 5	2.4067	8.2883	1.02E-09	3.6532	9.5241	8.54E-10	1.77E-02	-2.7425	-2.9703	-0.3414
20	203759_at	SIAT4C	sialyltransferase 4C (beta-galactoside alpha-2,3-sialyltransferase)	1.5035	8.2782	1.09E-09				6.50E-02	-2.0018	-0.9961	0.0342
21	240856_at		** no description **	1.7989	8.2080	1.67E-09	2.0481	7.7313	1.99E-06	2.82E-01	-1.1147	-0.6355	0.1982
22	226654_at	MUC12	mucin 12	3.0988	8.0394	4.66E-09	4.2406	7.1298	2.65E-05	4.95E-03	-3.3015	-3.8841	-0.8334
23	229499_at	CAPN13	calpain 13	1.2187	7.8466	1.49E-08	1.2837	6.4588	4.59E-04	5.49E-01	-0.6115	-0.6903	0.3801
24	206422_at	GCG	glucagon	3.5394	7.8128	1.82E-08	6.0957	7.7872	1.56E-06	5.68E-01	-0.5848	-0.9049	0.5168
25	236681_at	HOXD13	homeo box D13	1.4419	7.5188	1.03E-07	1.6533	6.3341	7.75E-04	2.01E-01	-1.3466	-0.6199	0.1437
26	221024_s_at	SLC2A10	solute carrier family 2 (facilitated glucose transporter), member 10	1.5552	7.4735	1.35E-07	1.6304	5.6695	1.20E-02	7.86E-01	-0.2784	-0.5100	0.3951
27	238862_at	DKFZp761N11	hypothetical protein DKFZp761N1114	1.3657	7.4657	1.41E-07	1.5027	7.1762	2.17E-05	2.42E-01	-1.2275	-0.2082	0.0577
28	201482_at	QSCN6	quiescin Q6	1.3243	7.4495	1.55E-07	1.4197	7.2690	1.46E-05	2.20E-01	-1.2733	-0.9080	0.2246
29	210103_s_at	FOXA2	forkhead box A2	1.3894	7.4289	1.75E-07	1.4913	6.2272	1.21E-03	1.13E-01	-1.6815	-0.9156	0.1081
30	213993_at	SPON1	spondin 1, extracellular matrix protein	1.4348	7.4099	1.95E-07	1.6082	6.6934	1.71E-04	1.19E-01	-1.6442	-0.7080	0.0878
31	209436_at	SPON1	spondin 1, extracellular matrix protein	1.5394	7.1992	6.59E-07	1.7567	6.6098	2.43E-04	1.11E-01	-1.6837	-1.5765	0.1771
32	234994_at	KIAA1913	KIAA1913	2.0243	7.1920	6.87E-07	2.3745	6.1586	1.61E-03	4.51E-02	-2.1685	-2.3949	-0.0299
33	204519_s_at	TM4SF11	transmembrane 4 superfamily member 11 (plasmolipin)	1.5123	7.1801	7.35E-07	1.7330	6.4681	4.42E-04	1.52E-02	-2.7824	-1.6258	-0.2071
34	213134_x_at	BTG3	BTG family, member 3	1.3761	7.1419	9.14E-07	1.4909	6.1257	1.85E-03	4.03E-01	-0.8587	-1.0225	0.4315
35	206070_s_at	EPHA3	EPH receptor A3	1.3440	7.0592	1.46E-06				7.16E-01	0.3698	-0.1398	0.1992
36	201889_at	FAM3C	family with sequence similarity 3, member C	1.5846	6.9954	2.10E-06	1.8871	7.1044	2.96E-05	1.77E-01	-1.4134	-2.1726	0.4361
37	239805_at	SLC13A2	solute carrier family 13 (sodium-dependent dicarboxylate transporter), member 2	1.4052	6.9691	2.43E-06				3.14E-01	-1.0401	-0.7317	0.2496
38	218187_s_at	FLJ20989	hypothetical protein FLJ20989	1.3131	6.9597	2.57E-06				2.67E-03	-3.5484	-1.9436	-0.4900
39	201798_s_at	FER1L3	fer-1-like 3, myoferlin (C. elegans)	1.4386	6.9150	3.30E-06	1.5077	5.8090	6.80E-03	6.52E-02	-1.9885	-2.4341	0.0839
40	207397_s_at	HOXD13	homeo box D13	1.2156	6.8953	3.68E-06	1.3278	5.4274	3.18E-02	3.01E-01	-1.0705	-0.1530	0.0507
41	205548_s_at	BTG3	BTG family, member 3	1.3727	6.8644	4.38E-06	1.4636	5.5270	2.13E-02	5.93E-01	-0.5445	-0.6543	0.3860
42	207080_s_at	PYY	peptide YY	2.9642	6.8281	5.36E-06	4.4363	6.1558	1.63E-03	8.57E-01	0.1831	-0.5225	0.6204
43	206104_at	ISL1	ISL1 transcription factor, LIM/homeodomain, (islet-1)	1.2491	6.7817	6.93E-06	1.3294	5.3926	3.65E-02	2.53E-01	-1.1876	-0.6539	0.1853
44	203961_at	NEBL	nebulette	1.5345	6.6278	1.62E-05	1.8643	7.7938	1.52E-06	2.30E-01	-1.2620	-1.2328	0.3265
45	208121_s_at	PTPRO	protein tyrosine phosphatase, receptor type, O	1.5772	6.6010	1.87E-05	1.7949	6.6295	2.23E-04	2.18E-01	1.2917	-0.0552	0.2220
46	236129_at	GALNT5	UDP-N-acetyl-alpha-D-galactosamine	1.3923	6.5855	2.04E-05	1.5111	6.1059	2.00E-03	2.44E-02	-2.4706	-0.5979	-0.0471
47	203698_s_at	FRZB	frizzled-related protein			2.08E-05	1.6958	7.1867	2.08E-05	2.48E-01	1.1964	-0.0771	0.2782
48	204351_at	S100P	S100 calcium binding protein P	2.5316	6.5625	2.31E-05	3.2208	6.0619	2.40E-03	4.68E-02	-2.1574	-3.6312	-0.0295
49	205042_at	GNE	glucosamine (UDP-N-acetyl)-2-epimerase/N-acetylmannosamine kinase	1.6163	6.4563	4.11E-05	2.0082	6.7357	1.43E-04	9.31E-03	-2.9329	-2.2337	-0.3643
50	205979_at	SCGB2A1	secretoglobin, family 2A, member 1	1.7328	6.4027	5.48E-05	2.0193	5.5811	1.72E-02	1.14E-01	-1.6938	-0.6383	0.0771
51	205927_s_at	CTSE	cathepsin E	1.4237	6.3675	6.62E-05	1.5846	6.0712	2.31E-03	5.49E-02	-2.0671	-1.2770	0.0147
52	229893_at	FRMD3	FERM domain containing 3	1.2730	6.3194	8.55E-05				1.83E-01	-1.3901	-1.1336	0.2342
53	228004_at	C20orf56	chromosome 20 open reading frame 56	1.7141	6.2459	1.26E-04				6.93E-01	-0.4040	-0.4126	0.2826
54	208450_at	LGALS2	lectin, galactoside-binding, soluble, 2 (galectin 2)	2.0310	6.2396	1.31E-04	2.4773	5.3780	3.87E-02	7.57E-02	-1.9311	-1.7705	0.0999
55	211253_x_at	PYY	peptide YY	1.3778	6.1703	1.88E-04	1.5825	5.5802	1.72E-02	3.08E-01	-1.0510	-0.7604	0.2555
56	228821_at	SIAT2	sialyltransferase 2 (monosialoganglioside sialyltransferase)	1.2800	6.1437	2.16E-04				1.80E-01	-1.4124	-0.1647	0.0341
57	214601_at	TPH1	tryptophan hydroxylase 1 (tryptophan 5-monooxygenase)	1.4092	6.0972	2.75E-04	1.6272	5.3527	4.27E-02	6.10E-01	0.5265	-0.1518	0.2462
58	213369_at	PCDH21	protocadherin 21	1.4794	6.0159	4.20E-04	1.7538	6.0814	2.22E-03	1.50E-01	-1.5266	-0.9169	0.1555
59	204686_at	IRS1	insulin receptor substrate 1	1.4809	6.0115	4.29E-04				2.47E-01	-1.2000	-1.1810	0.3247
60	202709_at	FMOD	fibromodulin	1.2559	5.9660	5.43E-04				9.98E-01	0.0024	-0.3258	0.3265
61	234709_at	CAPN13	calpain 13	1.2740	5.9574	5.67E-04	1.2837	5.5315	2.09E-02	2.69E-01	-1.1440	-0.4154	0.1239
62	218692_at	FLJ20366	hypothetical protein FLJ20366	1.2335	5.9139	7.08E-04				9.13E-01	-0.1113	-0.5029	0.4540
63	218532_s_at	FLJ20152	hypothetical protein FLJ20152	1.5696	5.8952	7.79E-04	1.8512	5.6880	1.11E-02	8.96E-02	-1.8034	-2.5468	0.2020
64	242414_at		** no description **	1.1722	5.8510	9.76E-04				9.69E-01	-0.0401	-0.2909	0.2801
65	212935_at	MCF2L	MCF.2 cell line derived transforming sequence-like	1.2007	5.8489	9.86E-04				6.83E-01	-0.4164	-0.5219	0.3506
66	218510_x_at	FLJ20152	hypothetical protein FLJ20152	1.4942	5.8115	1.19E-03	1.7263	5.4431	2.98E-02	1.77E-01	-1.4185	-2.5309	0.5086
67	213921_at	SST	somatostatin	1.7335	5.8030	1.24E-03				5.61E-01	-0.5941	-0.5395	0.3039
68	232321_at	MUC17	mucin 17	1.5373	5.7650	1.51E-03	1.6719	5.7561	8.44E-03	3.94E-02	-2.2843	-1.2222	-0.0353
69	205464_at	SCNN1B	sodium channel, nonvoltage-gated 1, beta (Liddle syndrome)	1.5884	5.7391	1.72E-03				3.00E-02	-2.3775	-2.0960	-0.1218
70	212098_at	LOC151162	hypothetical protein LOC151162	1.2162	5.7307	1.79E-03	1.3275	6.0706	2.32E-03	8.25E-02	-1.8581	-1.2610	0.0853
71	219973_at	FLJ23548	hypothetical protein FLJ23548	1.0946	5.6928	2.16E-03				2.63E-01	-1.1632	-0.0764	0.0225
72	203769_s_at	STS	steroid sulfatase (microsomal), arylsulfatase C, isozyme S	1.1896	5.6677	2.45E-03				6.13E-01	0.5151	-0.2235	0.3673
73	230645_at	FRMD3	FERM domain containing 3	1.2643	5.6646	2.49E-03				3.38E-01	-0.9874	-0.8492	0.3083
74	213432_at	MUC5B	mucin 5, subtype B, tracheobronchial			3.09E-03	2.3060	6.0011	3.09E-03	1.72E-01	-1.4427	-1.2975	0.2553
75	204781_s_at	FAS	Fas (TNF receptor superfamily member)	1.2457	5.5988	3.44E-03				9.60E-01	-0.0515	-0.5923	0.5646
76	203021_at	SLPI	secretory leukocyte protease inhibitor (antileukoproteinase)	1.6300	5.5982	3.46E-03	2.2457	7.0224	4.20E-05	9.88E-03	-2.9491	-3.1941	-0.5152
77	204044_at	QPRT	quinolinate phosphoribosyltransferase (nicotinate-nucleotide pyrophosphorylase (carboxylating))	1.2874	5.5770	3.84E-03				1.01E-01	-1.8025	-0.6642	0.0689
78	228256_s_at	EPB41L4A	erythrocyte membrane protein band 4.1 like 4A	1.2835	5.5607	4.15E-03				3.17E-01	-1.0324	-0.5961	0.2054
79	219033_at	PARP8	poly (ADP-ribose) polymerase family, member 8			4.48E-03	1.2434	5.9109	4.48E-03	8.29E-01	0.2199	-0.2334	0.2877
80	235004_at	RBM24	RNA binding motif protein 24	1.3389	5.5145	5.21E-03				2.33E-01	-1.2599	-0.5129	0.1390
81	205009_at	TFF1	trefoil factor 1 (breast cancer, estrogen-inducible sequence expressed in)	2.2026	5.5133	5.24E-03				8.36E-02	-1.8608	-2.7462	0.1932
82	212959_s_at	MGC4170	MGC4170 protein			5.56E-03	1.5719	5.8581	5.56E-03	2.94E-01	-1.0880	-1.8811	0.6093
83	213423_x_at	TUSC3	tumor suppressor candidate 3	1.4004	5.4510	7.09E-03				7.01E-01	-0.3902	-0.3242	0.2231
84	211719_x_at	FN1	fibronectin 1	1.8475	5.4506	7.11E-03				2.60E-01	1.1686	-0.8995	3.1093
85	213280_at	GARNL4	GTPase activating Rap/RanGAP domain-like 4	1.2152	5.4296	7.86E-03				4.51E-02	-2.1642	-0.8856	-0.0110
86	222258_s_at	SH3BP4	SH3-domain binding protein 4	1.2523	5.4281	7.92E-03	1.3838	5.6336	1.39E-02	7.24E-01	-0.3598	-0.7497	0.5342
87	205221_at	HGD	homogentisate 1,2-dioxygenase (homogentisate oxidase)	1.3595	5.4277	7.94E-03				1.74E-01	-1.4227	-0.8949	0.1761
88	226050_at	C13orf11	chromosome 13 open reading frame 11	1.2961	5.4095	8.67E-03				2.65E-01	-1.1581	-1.2866	0.3803
89	225591_at	FBXO25	F-box protein 25	1.1734	5.3977	9.18E-03				3.52E-01	-0.9692	-0.5157	0.1986
90	209228_x_at	TUSC3	tumor suppressor candidate 3	1.3320	5.3700	1.05E-02				2.11E-01	1.3517	-0.1411	0.5509
91	214798_at	KIAA0703	KIAA0703 gene product	1.2832	5.3679	1.06E-02				9.82E-01	0.0230	-0.5970	0.6096
92	212573_at	KIAA0830	KIAA0830 protein			1.09E-02	1.4028	5.6938	1.09E-02	2.89E-02	-2.4389	-3.0784	-0.1956
93	220136_s_at	CRYBA2	crystallin, beta A2	1.1975	5.3523	1.14E-02				5.55E-01	-0.6017	-0.3532	0.1966
94	41469_at	PI3	protease inhibitor 3, skin-derived (SKALP)	1.5984	5.3485	1.16E-02	2.1561	5.8717	5.26E-03	4.36E-02	-2.1967	-3.1937	-0.0531
95	210643_at	TNFSF11	tumor necrosis factor (ligand) superfamily, member 11	1.0847	5.3372	1.23E-02				9.40E-01	-0.0779	-0.2315	0.2159
96	203697_at	FRZB	frizzled-related protein			1.38E-02	1.6734	5.6350	1.38E-02	7.36E-01	0.3430	-0.3287	0.4560
97	205081_at	CRIP1	cysteine-rich protein 1 (intestinal)	1.4710	5.3107	1.39E-02	1.7786	5.5089	2.29E-02	9.96E-02	-1.7726	-2.4028	0.2364
98	212448_at	NEDD4L	neural precursor cell expressed, developmentally down-regulated 4-like	1.2048	5.3009	1.46E-02				1.90E-02	-2.5972	-1.0089	-0.1038
99	210495_x_at	FN1	fibronectin 1	1.7618	5.2865	1.56E-02				2.53E-01	1.1889	-0.7749	2.7368
100	212464_s_at	FN1	fibronectin 1	1.8202	5.2855	1.57E-02				2.72E-01	1.1408	-0.8472	2.8050
101	219734_at	SIDT1	SID1 transmembrane family, member 1	1.2674	5.2552	1.81E-02				5.73E-01	-0.5770	-0.4726	0.2719
102	227048_at	LAMA1	laminin, alpha 1			1.94E-02	1.9692	5.5506	1.94E-02	4.30E-02	-2.2108	-2.5885	-0.0476
103	216442_x_at	FN1	fibronectin 1	1.7670	5.2217	2.12E-02				2.67E-01	1.1493	-0.8418	2.8321
104	209437_s_at	SPON1	spondin 1, extracellular matrix protein	1.2281	5.2215	2.12E-02				4.36E-01	0.8127	-0.1527	0.3266
105	206502_s_at	INSM1	insulinoma-associated 1	1.2440	5.2145	2.19E-02	1.4613	5.5757	1.75E-02	5.49E-01	0.6123	-0.0582	0.1057
106	201097_s_at	ARF4	ADP-ribosylation factor 4	1.2820	5.2132	2.21E-02				1.56E-01	-1.5017	-2.7260	0.4863
107	203649_s_at	PLA2G2A	phospholipase A2, group IIA (platelets, synovial fluid)	1.9975	5.2082	2.26E-02				2.57E-01	-1.1727	-3.1818	0.9107
108	218976_at	DNAJC12	DnaJ (Hsp40) homolog, subfamily C, member 12	1.3074	5.2059	2.28E-02				4.86E-01	-0.7120	-0.2867	0.1421
109	218211_s_at	MLPH	melanophilin	1.3781	5.1857	2.51E-02				6.65E-01	-0.4411	-1.1438	0.7489
110	203962_s_at	NEBL	nebulette	1.4431	5.1725	2.67E-02	1.6869	5.5034	2.34E-02	2.88E-01	-1.1152	-0.8238	0.2690
111	229555_at	GALNT5	UDP-N-acetyl-alpha-D-galactosamine	1.1612	5.1681	2.72E-02				9.63E-01	0.0469	-0.4312	0.4503
112	237183_at	GALNT5	UDP-N-acetyl-alpha-D-galactosamine	1.1999	5.1605	2.82E-02				5.07E-01	-0.6779	-0.2433	0.1249
113	211864_s_at	FER1L3	fer-1-like 3, myoferlin (C. elegans)	1.3242	5.1576	2.86E-02				2.60E-01	-1.1717	-0.9109	0.2648
114	212186_at	ACACA	acetyl-Coenzyme A carboxylase alpha	1.1447	5.1422	3.07E-02				4.68E-01	0.7418	-0.2308	0.4809
115	239814_at		** no description **			3.21E-02	1.2166	5.4248	3.21E-02	5.39E-01	0.6303	-0.1514	0.2763
116	219909_at	MMP28	matrix metalloproteinase 28	1.2335	5.1262	3.31E-02				9.59E-02	-1.7646	-0.9624	0.0864
117	213308_at	SHANK2	SH3 and multiple ankyrin repeat domains 2	1.2366	5.1150	3.49E-02				6.25E-01	0.4985	-0.1723	0.2789
118	200677_at	PTTG1IP	pituitary tumor-transforming 1 interacting protein			3.52E-02	1.2472	5.4015	3.52E-02	6.80E-02	-1.9938	-1.9492	0.0799
119	221577_x_at	GDF15	growth differentiation factor 15	1.7442	5.1093	3.58E-02				1.17E-01	-1.6687	-0.7535	0.0942
120	205490_x_at	GJB3	gap junction protein, beta 3, 31kDa (connexin 31)	1.2239	5.0952	3.82E-02				9.12E-02	-1.8032	-1.1368	0.0942
121	231814_at	MUC11	mucin 11	1.7000	5.0934	3.86E-02	2.3413	5.4097	3.41E-02	1.48E-01	-1.5371	-0.5833	0.0979
122	205518_s_at	CMAH	cytidine monophosphate-N-acetylneuraminic acid hydroxylase (CMP-N-acetylneuraminate monooxygenase)	1.3496	5.0848	4.01E-02				6.00E-01	0.5344	-0.2306	0.3865
123	203691_at	PI3	protease inhibitor 3, skin-derived (SKALP)	1.7037	5.0784	4.13E-02	2.3708	5.4493	2.91E-02	7.84E-03	-3.0135	-3.5761	-0.6304
124	238378_at		** no description **	1.1627	5.0641	4.41E-02				9.94E-01	0.0083	-0.1082	0.1090
125	212570_at	KIAA0830	KIAA0830 protein			4.49E-02	1.2827	5.3405	4.49E-02	3.02E-01	-1.0690	-0.6966	0.2309
126	244553_at		** no description **	1.1397	5.0518	4.67E-02	1.2518	6.4213	5.37E-04	3.56E-01	-0.9510	-0.1984	0.0756

Figure 4.2: Genes and probesets elevated in distal tissues relative to proximal tissues.

D.2.3 RT-PCR validation of proximal-distal genes

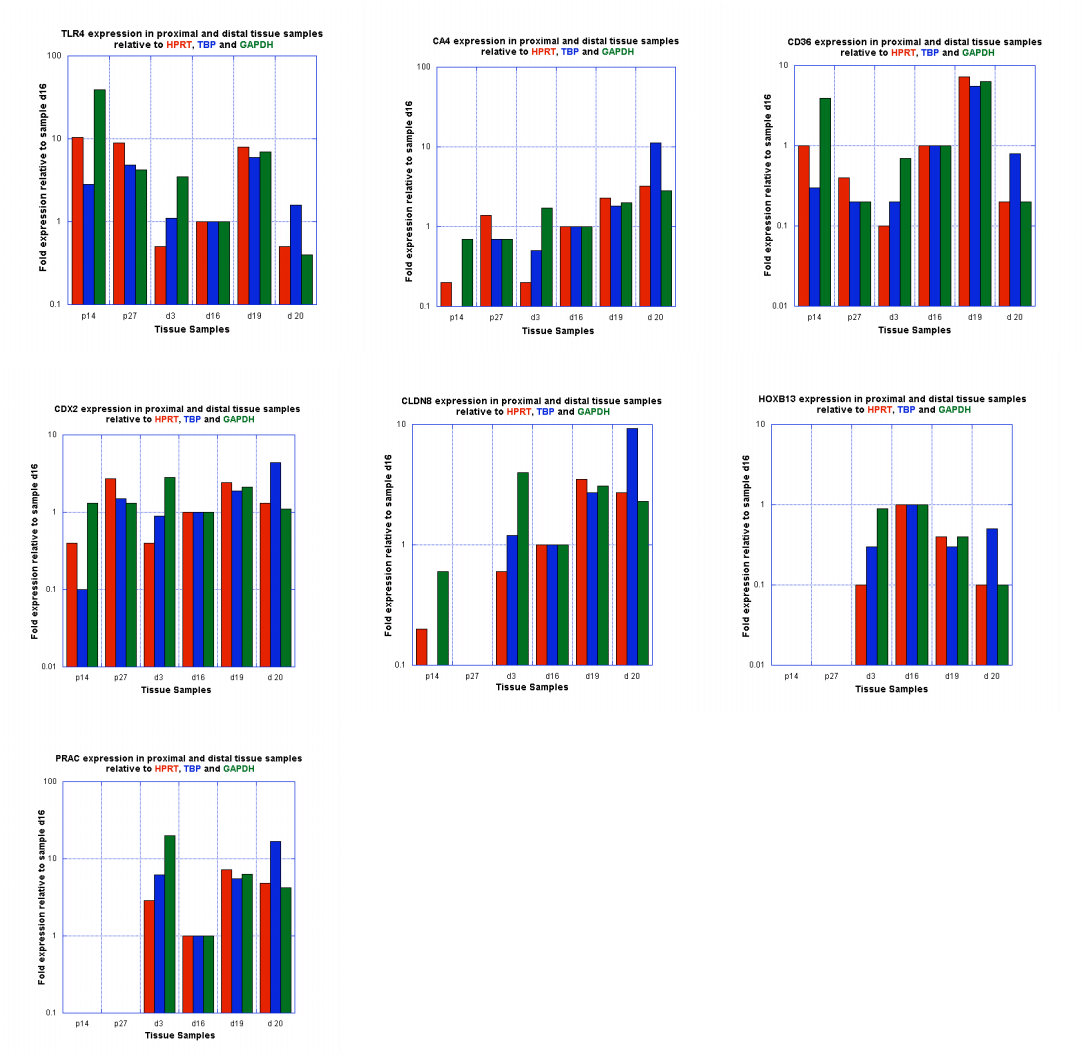


Figure 4.3: Results from TaqMan RT-PCR experiments measuring selected genes chosen for RT-PCR validation from those genes identified by microarray data to be differentially expressed in proximal vs. distal colorectal tissues. For this experiment, only non-diseased specimens were used.

D.3 Discovery - differential display

D.3.1 Annotation of differential display sequences

Table D.4: Annotation of Sequence IDs discovered by differential display

Sequence ID	Gene Symbols with Alignment
1	-NA-
2	GIF
3	NONE
4	C20orf199
5	C20orf199,TALDO1
6	-NA-
7	KIAA1199,MIRN549,FAM108C1,ELK4,SLC45A3,MFSD4,PARD3B,EXOC6
8	NRGN,VSIG2,C11orf61,ESAM
9	PVT1
10	MSH3
11	PROC,MAP3K2,MTMR2,MAML2,ASNSL1,MARCH6,ROPN1L,LOC728124
12	-NA-
13	KIAA1600
14	NCK2
15	KRTAP5-9,OR7E87P,NADSYN1,KRTAP5-8,KRTAP5-10,KRTAP5-7,etc
16	TTC28,PRPH,TROAP,SPATS2,FLJ13236,TUBA1C,C1QL4
17	CADPS
18	MPZL1,SAC
19	KHDRBS1,KPNA6,TMEM39B
20	ATQL3,GIMAP4,GIMAP7,GIMAP6
21	TG
22	APP
23	-NA-
24	CLIC2,VBP1,RAB39B,PHF10P1,LOC401622,LOC553939,TMLHE,etc
25	SLC7A1,KIAA0774,hCG_2020170
26	LRRC37A,KIAA1267,LOC644246
27	LPO,MPO,SUPT4H1,BZRAP1,RNF43,MIRN142,HSF5,Sept4,Tex14,etc
28	LPO,MPO,SUPT4H1,BZRAP1,RNF43,MIRN142,HSF5,PER2
29	LPO,MPO,SUPT4H1,BZRAP1,RNF43,MIRN142,HSF5,Sept4,Tex14,etc
30	GAPDHL16,LOC724105,MLH1,LRRFIP2
31	LOC750003
32	L1.1,PDHX,LOC440264
33	NONE
34	NONE
35	FAM135A
36	-NA-
37	MGC24039,hCG_1644239,MRPL30P2,LOC645619,PIN1L,LRRC7,etc
38	REG4
39	NONE
40	ZKSCAN1
41	CKSCAN1
42	RAB8A
43	GIF
44	NONE
45	DMBT1
46	ZNF800,Zfp800
47	PMS2L3
48	PTEN
49	DEFA6

Continued on Next Page...

Table D.4 – Continued

Sequence ID	Gene Symbols with Alignment
50	NONE
51	-NA-
52	MAGEF1,VPS8
53	HEPH
54	TFRC,SDHALP1
55	MLLT3,AF-9
56	-NA-
57	NONE
58	LIMA1
59	POF1B
60	GW112
61	NONE
62	S100P
63	NONE
64	-NA-
65	STAT2
66	NKCC1,SLC12A2
67	NONE
68	RPS4X
69	RPL14
70	RPL14
71	ASH1L
72	IFITM2,IFITM1
73	TERF1,KCNB2,RPESP,LOC286191
74	GNL3L
75	RHOQ
76	-NA-
77	EPSTI1
78	ABCB11,DHRS9
79	CCDC123
80	RPL7L1
81	PLCB4,C20orf103,PAK7,C12orf42,MNS1,TEX9,ZNF518A,BLNK
82	LASS6,NOSTRIN
83	PROS1,LGMN,GOLGA5,OR5BP1P,OR5AO1P,LIPI,C21orf126
84	IKBKAP,CTNNAL1,C9orf5,C9orf6,MIRN32
85	GLT8D1
86	FAM20B
87	KIF3C,ASXL2,LOC751599
88	MRPS36P2
89	MST157,MST153,SLC26A2,PDE6A,LOC644762
90	GSDML
91	COPS4,PLAC8
92	ATP9B,NFATC1
93	LOC460550,Col8a1,2610528E23Rik
94	ERGIC3
95	-NA-
96	SCARB2,STBD1,FLJ25770,GRIN2A,ZNF66,STAG1,ZNF204,PRSS16,etc
97	SELT
98	F3,ABCD3
99	MAP3K5,PEX7
100	LPO,MPO,SUPT4H1,BZRAP1,RNF43,MIRN142,HSF5
101	OVOL2,RPL15P1,CNIH4,WDR26,XRCC2
102	GALNT6
103	KIAA1199,MIRN549,CDKL1,ATP5S,L2HGDH
104	CS,MYL6,SMARCC2,USP52,RNF41,CNPY2,OBFC2B,COQ10A,SLC39A5,EXO1,WDR64
105	ZNF800

Continued on Next Page. . .

Table D.4 – Continued

Sequence ID	Gene Symbols with Alignment
106	CCNI
107	-NA-
108	SNTB2,TERF2,VPS4A,NIP7,PDF,CYB5B,COG8,TMED6
109	TRIO,FAM105A
110	SLC7A1,KIAA0774,hCG_2020170
111	ACACA
112	KRTAP5-9,OR7E87P,NADSYN1,KRTAP5-8,KRTAP5-10,KRTAP5-7,etc
113	ACACA
114	ALPK1,LAK
115	OCC-1
116	ARHGEF10,KBTBD11,MIRN596
117	GLG1,LOC440348,LOC440386,LOC497190,PDPR,PDXDC2,AARS,etc
118	CIR,SCRN3,GPR155,FLJ46347,ARFGEF1,CPA6
119	-NA-
120	FNTB,LOC389072
121	STARD3NL
122	ORC2L
123	STARD3NL
124	STARD3NL
125	PLCB4,C20orf103,PAK7,DLGAP4,C20orf24,TGIF2,SLA2,ANPEP,etc
126	MYO1E,LDHAL6B,PAK7
127	CPA6
128	SRPX2,SYTL4,Sytl4,Tspan6,Tnmd,Srpx2
129	FLVCR2,BATF,C14orf1,TTLL5,RPS24P2
130	PHF14
131	-NA-
132	POLR1A,MAOA,CNOT2,GSTM3,GSTM5,EPS8L3
133	C1orf123
134	CMIP,PLCG2
135	DYNC1LI2
136	FOXD4L3,FOXD4L5,CBWD5,FOXD4L4,RP11-460E7.5,IGKV1OR-3
137	UCK2
138	ERCC3,MAP3K2,CYP27C1,TBL1X,PSMA6,KIAA0391,PPP2R3C,etc
139	ESM1,SCAPER,DNAH14,RP11-328N1.1
140	UGCGL2
141	FBXW7
142	GNG4
143	SLC39A10,DNAH7
144	JMJD1C
145	JMJD1C
146	ARMCX6
147	-NA-
148	FOE,WAPAL
149	-NA-
150	-NA-
151	-NA-
152	GMDS
153	RPS7,C3orf67
154	HPF4, HTF1,ZNF85,TSHB,TSPAN2,CLGN,SCOC,LOC152586,ELMOD2,etc
155	CNGA3,VWA3B,PLDN,SQRDL,C15orf21,SESN1,C6orf182,C6orf183,etc
156	TCF12,LOC145783,CAP1,CEP152,EID1,SHC4,LOC724065
157	RPRM
158	PROC,MAP3K2,MTMR2,MAML2,ASNSL1,MARCH6,ROPN1L,LOC728124
159	FAM107A,FAM3D,C3orf67,PRPF3,KIAA0460,SEPHS1
160	-NA-
161	RHOBTB3,SPATA9,ZFP30,ZNF571,ZNF540,ZNF793

Continued on Next Page. . .

Table D.4 – Continued

Sequence ID	Gene Symbols with Alignment
162	ATAD1
163	NR2C2,ZFYVE20,MRPS25,OR7E15P,OR7E10P,OR7E8P,OR7E96P,etc
164	POF1B
165	GLRX5,TCL1B,TCL6,SNHG10
166	GLRX5
167	IL7,C18orf1,IGSF6,METTL9,OTOA
168	ZNF341
169	-NA-
170	AOF2
171	DNASE1L3,FLNB,ABHD6,INSR,DYNC1I2,NCAPD3,JAM3
172	EMP1,C12orf36
173	TRAPPC4
174	LRPPRC
175	PLEKHA5,RPL7P6
176	EHF
177	TEX261
178	GPB3
179	RPL7L1
180	SLC2A8,FAM129B,GARNL3,LRSAM1
181	RAB9A,EGFL6,MGC17403,LOC645769
182	SCRN1,FKBP14,PLEKHA8
183	PCDH21
184	CHMP44A
185	RPL7L1
186	DALRD3
187	PCDH21
188	LOC452328
189	KRTCAP2,LOC740337,KCP2
190	HMGB3
191	-NA-
192	OLA1
193	-NA-
194	APPL2
195	RPL13,DPEP1,CHMP1A,CDK10,C16orf7,CPNE7,ZNF276,SPATA2L,etc
196	G3BP2
197	HLA-DQA1,HLA-DQB1,HLA-DRB1
198	HPGD,GLRA3
199	-NA-
200	DYNC1LI2,XPC,LSM3,TMEM43,CHCHD4,TPRXL
201	PHACTR2
202	ROD1,Rod1
203	RPSAP14,LOC554203,FAM113B,CTRC,EFHD2,FHAD1,SPTLC3
204	SBDS
205	PTGES2,SLC25A25,LOC286208,C9orf119,LOC389791,MIRN199B,etc
206	-NA-
207	-NA-
208	LARP4
209	OGT
210	HN1L
211	TGFBI
212	MSTO1,ASH1L,MRPS29P1,LOC645676
213	-NA-
214	F3,ABCD3
215	OR7E15P,OR7E10P,OR7E8P,CLEC6A,OR7E149P,OR7E148P,OR7E140P,etc
216	CCNO,DHX29,SKIV2L2
217	CADPS

Continued on Next Page. . .

Table D.4 – Continued

Sequence ID	Gene Symbols with Alignment
218	KIFAP3,MRPS10P1
219	OR7E96P,FAM90A11,FAM90A24,FAM90A12
220	HDGF,SH2D2A,NES,C1orf66,BCAN,MRPL24,ISG20L2
221	RPS26L
222	SPINK4
223	PLAGL2
224	CALR
225	PFDN5,TAF1A,C1orf80,KIAA1822L,MIA3
226	-NA-
227	NDFG1,DRG1,NDRG1
228	TCP1
229	TM7SF3,FGFR1OP2
230	MUC13,ITGB5,HEG1,DRCC1
231	-NA-
232	MLL3
233	IFITM2
234	LOC462344,GNB2L1
235	CD164
236	ARFGEF2
237	DYNLRB1,HSPC162,BLP
238	-NA-
239	PIN1,UBE2L4,FBXL12,UBL5,OLFM2,LOC162993
240	BLCAP,BC10
241	-NA-
242	WARS
243	SPPP1
244	APP
245	SYF2
246	NQO1
247	ECOP
248	LOC739695,CPY2S1
249	-NA-
250	-NA-
251	-NA-
252	SLK,COL17A1,KIAA0204
253	ATP8,COX1,COX2,COX3,ND1,ND2,ND4,ND4L,ND6,CYTB,ND3
254	ATP6,ATP8,COX1,COX2,ND1,ND2,ND3,ND4L,ND6,ND4
255	RPL6
256	-NA-
257	HSP90AA1
258	DC24
259	BPHL
260	NCK2,Nck2
261	-NA-
262	ERAL1
263	-NA-
264	-NA-
265	WDR61
266	SERF2
267	BEST1,FTH1
268	PPP1R11,ZNRD1,RNF39
269	LDHB
270	CTSC
271	PPM1G
272	RETNLB
273	TACC2

Continued on Next Page...

Table D.4 – Continued

Sequence ID	Gene Symbols with Alignment
274	TACC2
275	-NA-
276	PBX3
277	RBMS1,ITGB6,C2orf12
278	-NA-
279	PLA2G2A,OTUD3,PLA2G2E
280	C9orf57,RP11-61E5.1
281	REG4
282	-NA-
283	KCNQ1,KCNQ1OT1
284	SOD1
285	MBP-1
286	-NA-
287	ENO1
288	MBP-1
289	GPRC5A
290	ATP10B
291	MUC12
292	SDCCAG1
293	KHDRBS1,KPNA6,TMEM39B
294	NUBP1
295	FAT
296	APEX1
297	GMEB1
298	SF3B1
299	PRDX1
300	-NA-
301	POMP
302	S100A11
303	OSBPL8
304	ITGA6
305	LOC749201
306	KIAA1370,EEF1B1,MYO5A,ARPP-19
307	-NA-
308	CLCA1
309	SLC12A2
310	CLCA4,CLCA1
311	GPSM3,AGER,NOTCH4,PBX2,RNF5,PPT2,AGPAT1,PRRT1,EGFL8
312	HMGB1
313	RNF130
314	ZNF263
315	VAMP3
316	KIAA1199,FAM108C1,MIRN549
317	TM9SF1
318	-NA-
319	-NA-
320	TCTP,TPT1
321	PIGR
322	GORASP1,WDR48,TTC21A
323	SNAPC1,HIF1A
324	ZNF223
325	S100A6
326	-NA-
327	BCAS1,SUMO1P1
328	-NA-
329	MYO5B,KIAA1119,ACAA2,SCARNA17

Continued on Next Page. . .

Table D.4 – Continued

Sequence ID	Gene Symbols with Alignment
330	CLCA1
331	HSM-2
332	VAT1,BRCA1,IFI35,RND2,RPL21P4
333	TFF2
334	-NA-
335	ATP8,COX1,COX2,COX3,ND1,ND2,ND4,ND4L,ND6,CYTB,ND3,ATP6
336	-NA-
337	KIAA1045,DNAJB5,LOC158383,GLULP,KRT8
338	REG

D.4 Discovery - GeneLogic microarray data

D.4.1 QC: Principal component plots

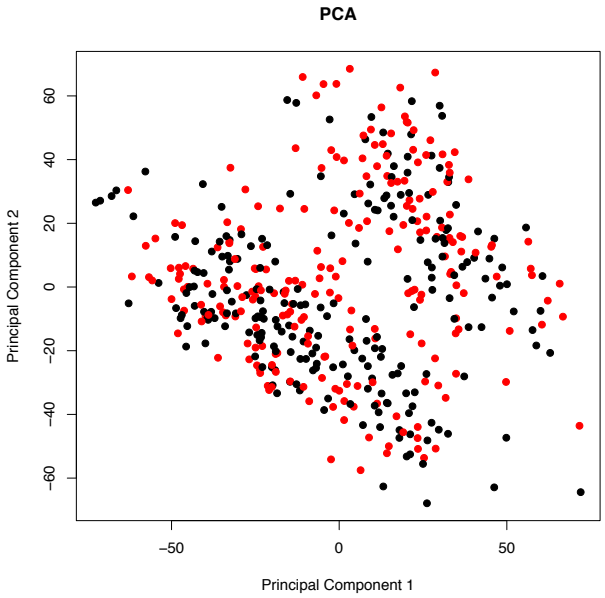


Figure 4.4: PCA analysis of GeneLogic data by gender. Tissues are indicated as male (black) or female (red). No effect is observed.

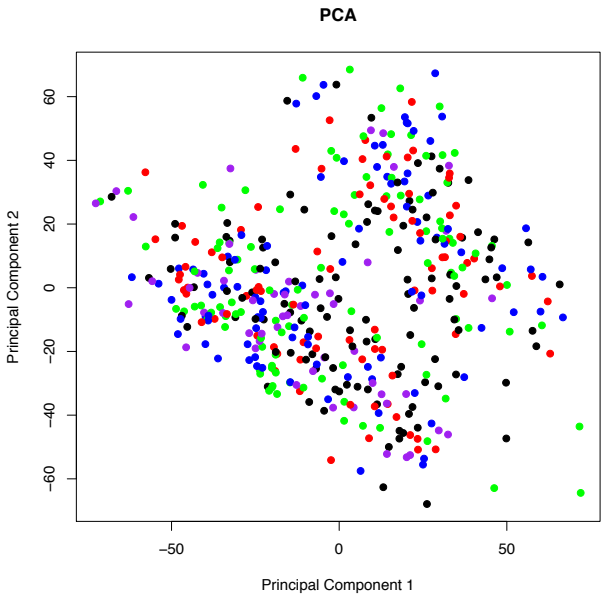


Figure 4.5: PCA analysis of GeneLogic data by age. No effect is observed.

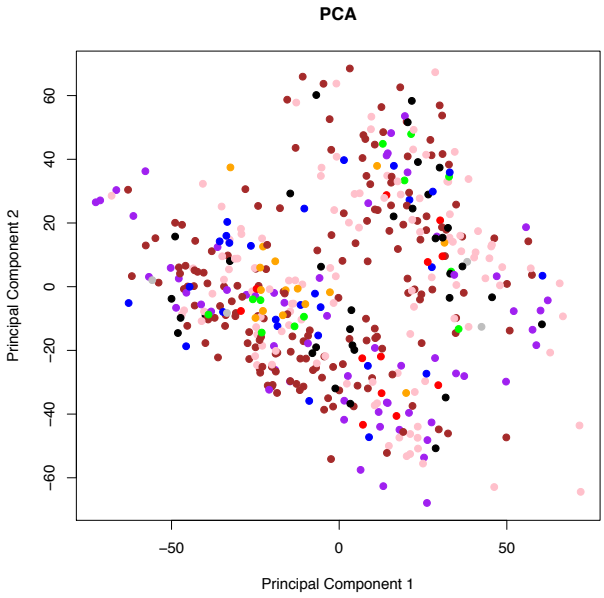


Figure 4.6: PCA analysis of GeneLogic data by GeneChip lot. No effect is observed.

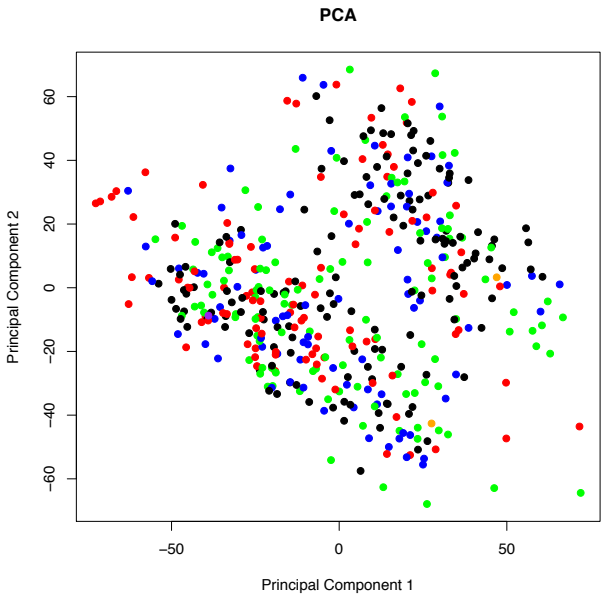


Figure 4.7: PCA analysis of GeneLogic data by operator. No effect is observed.

D.4.2 Probesets upregulated in neoplastic tissues

Table D.5: Top 108 differentially expressed probesets measured in GeneLogic data between neoplastic tissues and non-neoplastic controls

ProbeSetID	Symbol	Fold-Δ Log2	t statistic	P value (Bonf. corr.)	Likelihood
227475_at	FOXQ1	2.55	29.75	3.6210e-104	237.52
203256_at	CDH3	1.44	28.55	7.5028e-99	225.37
212942_s_at	KIAA1199	1.95	26.80	5.3906e-91	207.41
201506_at	TGFBI	1.68	26.48	1.5368e-89	204.09
204702_s_at	NFE2L3	1.02	26.29	1.0803e-88	202.15
201341_at	ENC1	1.07	26.27	1.2875e-88	201.98
219911_s_at	SLCO4A1	1.18	24.67	2.6996e-81	185.23
228754_at	SLC6A6	1.63	23.56	3.4022e-76	173.56
229215_at	ASCL2	1.69	23.44	1.3102e-75	172.22
201416_at	SOX4	1.30	22.93	2.7135e-73	166.92
227140_at	-NA-	2.46	22.85	6.8812e-73	165.99
203878_s_at	MMP11	1.19	22.66	4.9125e-72	164.04
218507_at	HIG2	1.01	22.55	1.5729e-71	162.88
222549_at	CLDN1	1.40	22.55	1.6596e-71	162.83
203962_s_at	NEBL	1.74	22.53	2.0269e-71	162.63
210511_s_at	INHBA	1.86	22.47	3.8857e-71	161.99
203510_at	MET	1.38	21.76	7.5862e-68	154.45
225806_at	JUB	1.00	21.62	3.4366e-67	152.95
203961_at	NEBL	1.46	21.58	4.9603e-67	152.59
201417_at	SOX4	1.52	21.54	8.1551e-67	152.09
218872_at	TESC	1.18	21.43	2.6138e-66	150.94
232151_at	7A5	1.01	20.90	7.7264e-64	145.28
225520_at	MTHFD1L	1.13	20.68	8.2218e-63	142.93
205983_at	DPEP1	1.55	20.67	8.5310e-63	142.89
200660_at	S100A11	1.05	20.31	4.0914e-61	139.05
224915_x_at	C20orf199	1.01	20.06	6.1077e-60	136.36
218704_at	RNF43	1.47	20.04	7.3682e-60	136.17
204259_at	MMP7	2.16	19.95	1.8796e-59	135.24
208712_at	CCND1	1.05	19.82	7.6509e-59	133.85
202936_s_at	SOX9	1.40	19.78	1.1471e-58	133.44
210766_s_at	CSE1L	1.05	19.67	4.0025e-58	132.20
218984_at	PUS7	1.02	19.58	9.9248e-58	131.30
221577_x_at	GDF15	2.04	19.58	1.0370e-57	131.26
219787_s_at	ECT2	1.27	19.52	1.9497e-57	130.63
226835_s_at	C20orf199	1.08	19.44	4.3228e-57	129.84
238021_s_at	hCG_1815491	1.45	19.26	2.8916e-56	127.95
206286_s_at	TDGF1	1.24	19.14	1.0484e-55	126.67
212070_at	GPR56	1.05	19.00	4.9946e-55	125.12
201563_at	SORD	1.30	18.79	4.3463e-54	122.96
225295_at	SLC39A10	1.01	18.73	8.7737e-54	122.27
213880_at	LGR5	1.96	18.65	1.8640e-53	121.52
222449_at	TMEPAI	1.28	18.59	3.8508e-53	120.80
225681_at	CTHRC1	2.54	18.58	4.2382e-53	120.70
207850_at	CXCL3	1.61	18.43	1.9547e-52	119.18
202954_at	UBE2C	1.03	18.29	8.8434e-52	117.68
202504_at	TRIM29	1.16	18.27	1.0582e-51	117.50
201666_at	TIMP1	1.30	18.22	1.9337e-51	116.90
37892_at	COL11A1	1.75	17.76	2.2548e-49	112.17
201195_s_at	SLC7A5	1.01	17.73	3.1032e-49	111.86
222696_at	AXIN2	1.39	17.48	4.5659e-48	109.18
210052_s_at	TPX2	1.00	17.37	1.4930e-47	108.01
204404_at	SLC12A2	1.21	17.19	9.8803e-47	106.13

Continued on Next Page...

Table D.5 – Continued

ProbeSetID	Symbol	Fold-Δ Log2	t statistic	P value (Bonf. corr.)	Likelihood
202935_s_at	SOX9	1.67	16.90	1.9687e-45	103.15
209875_s_at	SPP1	2.32	16.89	2.1288e-45	103.08
200832_s_at	SCD	1.44	16.36	5.3555e-43	97.58
204855_at	SERPINB5	1.79	16.35	5.9423e-43	97.48
202859_x_at	IL8	2.09	16.12	6.1468e-42	95.16
202431_s_at	MYC	1.03	15.72	3.7152e-40	91.08
222608_s_at	ANLN	1.10	15.57	1.7527e-39	89.54
212353_at	SULF1	1.44	15.44	6.9831e-39	88.17
202286_s_at	TACSTD2	1.72	15.29	3.1657e-38	86.67
209369_at	ANXA3	1.11	15.28	3.5129e-38	86.56
218963_s_at	KRT23	1.56	15.21	6.5586e-38	85.94
241031_at	FAM148A	1.04	15.20	7.8241e-38	85.77
206224_at	CST1	1.25	15.17	1.0704e-37	85.46
225541_at	RPL22L1	1.10	14.90	1.5487e-36	82.80
212190_at	SERPINE2	1.04	14.84	2.9064e-36	82.18
204170_s_at	CKS2	1.18	14.38	2.7401e-34	77.66
218796_at	FERMT1	1.02	14.36	3.3769e-34	77.45
212281_s_at	TMEM97	1.11	14.35	3.7198e-34	77.36
213905_x_at	BGN	1.15	14.33	4.4831e-34	77.17
209309_at	AZGP1	1.05	14.11	4.0182e-33	74.99
60474_at	FERMT1	1.05	14.04	8.1924e-33	74.29
205890_s_at	UBD	1.53	13.84	6.0300e-32	72.31
204051_s_at	SFRP4	1.39	13.80	8.5844e-32	71.95
204475_at	MMP1	2.18	13.78	1.0136e-31	71.79
202404_s_at	COL1A2	1.84	13.73	1.6858e-31	71.28
212354_at	SULF1	1.40	13.72	1.8127e-31	71.21
209955_s_at	FAP	1.01	13.70	2.3947e-31	70.94
202311_s_at	COL1A1	1.59	13.69	2.4402e-31	70.92
212344_at	SULF1	1.06	13.56	8.9299e-31	69.63
217996_at	PHLDA1	1.20	13.49	1.7597e-30	68.96
204470_at	CXCL1	1.25	13.36	6.1259e-30	67.72
224428_s_at	CDCA7	1.15	13.31	1.0181e-29	67.21
207457_s_at	LY6G6D	1.06	13.20	2.9857e-29	66.15
203083_at	THBS2	1.25	13.13	5.7208e-29	65.50
227174_at	WDR72	1.31	13.10	7.4420e-29	65.24
223062_s_at	PSAT1	1.16	13.09	8.0032e-29	65.17
205828_at	MMP3	1.45	13.01	1.7287e-28	64.40
226237_at	COL8A1	1.32	12.82	1.0645e-27	62.60
209218_at	SQLE	1.01	12.66	5.1032e-27	61.04
211506_s_at	IL8	1.54	12.42	4.6703e-26	58.85
205513_at	TCN1	1.17	12.41	5.0761e-26	58.77
204351_at	S100P	1.62	12.30	1.4154e-25	57.75
205476_at	CCL20	1.41	11.96	3.3567e-24	54.61
202310_s_at	COL1A1	1.47	11.65	5.3689e-23	51.86
209774_x_at	CXCL2	1.20	11.61	8.2683e-23	51.43
225835_at	SLC12A2	1.05	11.28	1.5011e-21	48.56
232252_at	DUSP27	1.13	11.07	1.0180e-20	46.67
204885_s_at	MSLN	1.04	10.86	6.1784e-20	44.88
212531_at	LCN2	1.61	10.70	2.4737e-19	43.51
207173_x_at	CDH11	1.07	10.60	5.6868e-19	42.68
225664_at	COL12A1	1.03	9.73	8.5473e-16	35.45
204580_at	MMP12	1.14	9.11	1.2174e-13	30.55
214974_x_at	CXCL5	1.10	8.50	1.2551e-11	25.98
209752_at	REG1A	1.80	8.25	7.7956e-11	24.18
205886_at	REG1B	1.37	7.99	5.1249e-10	22.32
205815_at	REG3A	1.18	6.43	1.4238e-05	12.29

Continued on Next Page...

Table D.5 – Continued

ProbeSetID	Symbol	Fold-Δ Log2	t statistic	P value (Bonf. corr.)	Likelihood

D.4.3 Probesets downregulated in neoplastic tissues

Table D.6: Top 338 down regulated probesets measured in GeneLogic data between neoplastic tissues and non-neoplastic controls

ProbeSetID	Symbol	Fold-Δ Log2	t statistic	P value (Bonf. corr.)	Likelihood
209612_s_at	ADH1B	−2.59	−26.92	1.5627e-91	208.64
229839_at	SCARA5	−1.57	−26.56	6.4307e-90	204.95
204719_at	ABCA8	−1.98	−25.89	7.3415e-87	197.96
203908_at	SLC4A4	−2.73	−25.75	2.9648e-86	196.57
226333_at	-NA-	−1.17	−25.19	1.1431e-83	190.66
209613_s_at	ADH1B	−2.56	−25.12	2.3664e-83	189.93
202242_at	TSPAN7	−1.47	−24.95	1.3849e-82	188.18
205200_at	CLEC3B	−1.84	−24.90	2.3655e-82	187.65
206209_s_at	CA4	−2.87	−24.74	1.2488e-81	185.99
224836_at	TP53INP2	−1.22	−24.24	2.5376e-79	180.71
226492_at	SEMA6D	−1.39	−23.98	3.9126e-78	178.00
206208_at	CA4	−2.29	−23.76	4.2398e-77	175.63
235849_at	SCARA5	−1.11	−23.74	5.1787e-77	175.43
203000_at	STMN2	−1.40	−23.73	5.6283e-77	175.35
230788_at	GCNT2	−1.18	−23.52	5.1534e-76	173.15
207761_s_at	METTL7A	−1.51	−23.37	2.5629e-75	171.55
209687_at	CXCL12	−2.37	−23.17	2.2443e-74	169.40
207003_at	GUCA1B	−2.65	−23.08	5.5636e-74	168.49
215118_s_at	IGHA1	−2.12	−22.97	1.9645e-73	167.24
204036_at	EDG2	−1.03	−22.94	2.5144e-73	166.99
228885_at	MAMDC2	−1.77	−22.60	9.2330e-72	163.41
205950_s_at	CA1	−3.23	−22.55	1.7343e-71	162.79
205382_s_at	CFD	−1.94	−22.37	1.1597e-70	160.90
207502_at	GUCA2B	−2.00	−22.26	3.6449e-70	159.76
230087_at	PRIMA1	−1.23	−22.24	4.6490e-70	159.52
211548_s_at	HPGD	−2.09	−22.12	1.7029e-69	158.23
205480_s_at	UGP2	−1.07	−22.11	1.8811e-69	158.13
220026_at	CLCA4	−3.50	−21.88	2.1103e-68	155.73
208399_s_at	EDN3	−1.28	−21.77	7.0540e-68	154.53
209301_at	CA2	−3.22	−21.70	1.5172e-67	153.77
203914_x_at	HPGD	−1.85	−21.61	3.7983e-67	152.85
223551_at	PKIB	−2.55	−21.43	2.7237e-66	150.90
201540_at	FHL1	−1.90	−21.39	4.1281e-66	150.48
209074_s_at	FAM107A	−1.20	−21.27	1.4586e-65	149.23
225207_at	PDK4	−2.23	−21.23	2.1872e-65	148.82
206637_at	P2RY14	−1.49	−21.08	1.1057e-64	147.21
202350_s_at	MATN2	−1.60	−21.08	1.1485e-64	147.18
228195_at	MGC13057	−1.36	−20.98	3.1460e-64	146.17
224480_s_at	AGPAT9	−1.22	−20.98	3.3588e-64	146.11
203913_s_at	HPGD	−2.08	−20.95	4.4763e-64	145.82
214696_at	C17orf91	−1.29	−20.92	5.9931e-64	145.53
213451_x_at	TNXB	−1.02	−20.89	8.0832e-64	145.24
204931_at	TCF21	−1.09	−20.77	2.9117e-63	143.96
230830_at	OSTbeta	−1.45	−20.75	3.7628e-63	143.71
219799_s_at	DHRS9	−1.56	−20.67	9.0095e-63	142.84

Continued on Next Page...

Table D.6 – Continued

ProbeSetID	Symbol	Fold-Δ Log2	t statistic	P value (Bonf. corr.)	Likelihood
209357_at	CITED2	−1.20	−20.47	7.1393e-62	140.78
204955_at	SRPX	−1.71	−20.45	9.6154e-62	140.49
210946_at	PPAP2A	−1.14	−20.43	1.1550e-61	140.30
224009_x_at	DHRS9	−1.96	−20.40	1.6235e-61	139.97
209735_at	ABCG2	−1.77	−20.33	3.3201e-61	139.25
223754_at	MGC13057	−1.01	−20.32	3.8011e-61	139.12
223395_at	ABI3BP	−1.66	−20.28	5.4705e-61	138.76
217546_at	MT1M	−2.35	−19.97	1.5872e-59	135.41
204834_at	FGL2	−1.24	−19.87	4.6910e-59	134.33
228469_at	PPID	−1.01	−19.82	7.6498e-59	133.85
223952_x_at	DHRS9	−1.80	−19.78	1.2076e-58	133.39
228504_at	-NA-	−1.74	−19.66	4.4381e-58	132.10
206422_at	GCG	−2.61	−19.64	5.1177e-58	131.96
206134_at	ADAMDEC1	−2.50	−19.62	6.4015e-58	131.74
242317_at	HIGD1A	−1.33	−19.56	1.2271e-57	131.09
203001_s_at	STMN2	−1.07	−19.56	1.2376e-57	131.08
205593_s_at	PDE9A	−1.34	−19.54	1.4556e-57	130.92
220834_at	MS4A12	−3.07	−19.47	3.2766e-57	130.11
205112_at	PLCE1	−1.01	−19.32	1.5203e-56	128.59
206149_at	CHP2	−1.96	−19.32	1.6551e-56	128.50
213624_at	SMPDL3A	−1.45	−19.30	2.0139e-56	128.31
220266_s_at	KLF4	−1.52	−19.25	3.4165e-56	127.78
227265_at	FGL2	−1.68	−19.22	4.5341e-56	127.50
222722_at	OGN	−1.84	−19.22	4.8108e-56	127.44
228707_at	CLDN23	−1.88	−19.20	5.4230e-56	127.32
205259_at	NR3C2	−1.43	−19.19	6.4136e-56	127.16
206710_s_at	EPB41L3	−1.12	−19.11	1.5034e-55	126.31
221841_s_at	KLF4	−1.63	−19.02	4.0331e-55	125.33
206641_at	TNFRSF17	−1.47	−18.95	8.0260e-55	124.64
213068_at	DPT	−1.89	−18.90	1.4438e-54	124.06
218756_s_at	MGC4172	−1.81	−18.89	1.5393e-54	124.00
214142_at	ZG16	−3.40	−18.69	1.2690e-53	121.90
204697_s_at	CHGA	−1.48	−18.65	1.9444e-53	121.48
201427_s_at	SEPP1	−1.33	−18.64	2.1326e-53	121.38
205464_at	SCNN1B	−1.28	−18.61	3.0900e-53	121.01
206377_at	FOXF2	−1.09	−18.59	3.6885e-53	120.84
206784_at	AQP8	−2.83	−18.54	6.0751e-53	120.34
227826_s_at	SORBS2	−2.42	−18.52	8.0799e-53	120.06
212814_at	AHCYL2	−1.24	−18.46	1.4643e-52	119.47
202037_s_at	SFRP1	−1.12	−18.31	7.3499e-52	117.86
225575_at	LIFR	−1.09	−18.26	1.2623e-51	117.33
221896_s_at	HIGD1A	−1.00	−18.08	8.0299e-51	115.49
215299_x_at	SULT1A1	−1.05	−17.98	2.2729e-50	114.45
222162_s_at	ADAMTS1	−1.35	−17.83	1.1434e-49	112.85
233565_s_at	SDCBP2	−1.00	−17.81	1.3509e-49	112.68
206143_at	SLC26A3	−3.46	−17.79	1.7751e-49	112.41
239272_at	MMP28	−1.01	−17.78	1.8703e-49	112.36
231925_at	P2RY1	−1.09	−17.71	3.9692e-49	111.61
219059_s_at	LYVE1	−1.01	−17.71	3.9979e-49	111.60
207980_s_at	CITED2	−1.06	−17.62	9.8867e-49	110.70
227827_at	SORBS2	−2.38	−17.60	1.2648e-48	110.46
206561_s_at	AKR1B10	−2.25	−17.60	1.2650e-48	110.46
204389_at	MAOA	−1.08	−17.58	1.5159e-48	110.28
208763_s_at	TSC22D3	−1.17	−17.39	1.1731e-47	108.24
209170_s_at	GPM6B	−1.11	−17.32	2.3624e-47	107.55
220376_at	LRRC19	−1.38	−17.29	3.4314e-47	107.18

Continued on Next Page...

Table D.6 – Continued

ProbeSetID	Symbol	Fold-Δ Log2	t statistic	P value (Bonf. corr.)	Likelihood
223623_at	C2orf40	−1.19	−17.26	4.6703e-47	106.87
231773_at	ANGPTL1	−1.22	−17.23	6.4784e-47	106.55
207080_s_at	PYY	−2.24	−17.22	6.8665e-47	106.49
202741_at	PRKACB	−1.09	−17.18	1.0092e-46	106.11
209763_at	CHRDL1	−1.44	−17.11	2.2821e-46	105.30
209373_at	MALL	−1.34	−17.10	2.3131e-46	105.28
218546_at	C1orf115	−1.07	−17.08	2.8693e-46	105.07
207432_at	BEST2	−1.16	−17.07	3.4404e-46	104.89
210299_s_at	FHL1	−1.64	−17.00	6.8604e-46	104.20
212859_x_at	MT1E	−1.30	−16.97	9.1429e-46	103.92
235146_at	-NA-	−1.24	−16.97	9.7659e-46	103.85
226303_at	PGM5	−1.92	−16.89	2.1891e-45	103.05
205554_s_at	DNASE1L3	−1.17	−16.86	2.9165e-45	102.76
229070_at	C6orf105	−1.94	−16.85	3.1379e-45	102.69
204388_s_at	MAOA	−1.03	−16.83	3.9010e-45	102.47
209167_at	GPM6B	−1.22	−16.75	9.0748e-45	101.63
204818_at	HSD17B2	−1.70	−16.71	1.3580e-44	101.23
206198_s_at	CEACAM7	−2.68	−16.71	1.3829e-44	101.22
221004_s_at	ITM2C	−1.12	−16.70	1.6186e-44	101.06
236300_at	-NA-	−1.10	−16.68	1.8924e-44	100.90
202746_at	ITM2A	−1.31	−16.61	3.9504e-44	100.17
226594_at	ENTPD5	−1.30	−16.41	3.2343e-43	98.08
206262_at	ADH1C	−1.95	−16.27	1.3136e-42	96.69
209791_at	PADI2	−1.27	−16.26	1.5481e-42	96.53
226430_at	RELL1	−1.01	−16.20	2.7752e-42	95.95
201739_at	SGK1	−1.42	−16.20	2.9020e-42	95.90
228961_at	MIER3	−1.11	−16.16	4.1773e-42	95.54
210298_x_at	FHL1	−1.63	−16.09	8.4499e-42	94.84
228706_s_at	CLDN23	−1.11	−16.04	1.5021e-41	94.27
205403_at	IL1R2	−1.44	−16.03	1.6136e-41	94.20
231120_x_at	PKIB	−1.32	−16.01	1.9491e-41	94.01
211848_s_at	CEACAM7	−2.15	−15.97	2.8448e-41	93.63
219014_at	PLAC8	−1.83	−15.96	3.2230e-41	93.51
227662_at	SYNPO2	−2.15	−15.93	4.4243e-41	93.20
201348_at	GPX3	−1.44	−15.91	5.4737e-41	92.98
226811_at	FAM46C	−1.12	−15.91	5.5719e-41	92.97
238143_at	LOC646627	−1.54	−15.89	6.8219e-41	92.77
212741_at	MAOA	−1.13	−15.88	7.7398e-41	92.64
201497_x_at	MYH11	−1.98	−15.87	8.0216e-41	92.60
217967_s_at	FAM129A	−1.68	−15.86	8.8420e-41	92.51
209667_at	CES2	−1.23	−15.83	1.2431e-40	92.17
207961_x_at	MYH11	−1.69	−15.78	2.1238e-40	91.64
213071_at	DPT	−1.26	−15.74	2.9628e-40	91.31
202920_at	ANK2	−1.10	−15.71	4.2079e-40	90.96
219669_at	CD177	−1.59	−15.70	4.7139e-40	90.85
206461_x_at	MT1H	−1.16	−15.68	5.4742e-40	90.70
215657_at	SLC26A3	−1.13	−15.68	6.0228e-40	90.60
203343_at	UGDH	−1.13	−15.66	6.7510e-40	90.49
211549_s_at	HPGD	−1.02	−15.65	7.7736e-40	90.35
206385_s_at	ANK3	−1.10	−15.64	8.9477e-40	90.21
212288_at	FNBP1	−1.05	−15.62	1.0929e-39	90.01
202992_at	C7	−1.14	−15.59	1.4140e-39	89.75
207977_s_at	DPT	−1.56	−15.56	1.9827e-39	89.42
217165_x_at	MT1F	−1.20	−15.55	2.1333e-39	89.35
225275_at	EDIL3	−1.15	−15.51	3.1588e-39	88.96
204745_x_at	MT1G	−1.23	−15.48	4.3947e-39	88.63

Continued on Next Page...

Table D.6 – Continued

ProbeSetID	Symbol	Fold-Δ Log2	t statistic	P value (Bonf. corr.)	Likelihood
228854_at	-NA-	−1.53	−15.36	1.5373e-38	87.38
236313_at	CDKN2B	−1.34	−15.31	2.3967e-38	86.94
201539_s_at	FHL1	−1.60	−15.30	2.6289e-38	86.85
224412_s_at	TRPM6	−1.44	−15.22	6.0251e-38	86.03
224959_at	SLC26A2	−2.19	−15.18	9.0545e-38	85.62
214091_s_at	GPX3	−1.33	−15.14	1.3327e-37	85.24
224963_at	SLC26A2	−1.92	−15.14	1.4366e-37	85.16
202731_at	PDCD4	−1.12	−15.10	2.1529e-37	84.76
202742_s_at	PRKACB	−1.09	−15.05	3.4779e-37	84.28
214505_s_at	FHL1	−1.32	−15.04	3.8786e-37	84.18
220812_s_at	HHLA2	−1.05	−15.03	4.3941e-37	84.05
220037_s_at	LYVE1	−1.13	−14.98	6.9533e-37	83.60
208581_x_at	MT1X	−1.13	−14.98	7.1423e-37	83.57
201496_x_at	MYH11	−2.64	−14.90	1.5144e-36	82.82
213629_x_at	MT1F	−1.24	−14.90	1.6107e-36	82.76
222717_at	SDPR	−1.19	−14.83	3.2225e-36	82.07
225720_at	SYNPO2	−1.80	−14.79	4.7454e-36	81.69
203766_s_at	LMOD1	−1.62	−14.76	6.7337e-36	81.34
204130_at	HSD11B2	−1.61	−14.74	8.1414e-36	81.15
225895_at	SYNPO2	−2.28	−14.72	9.6162e-36	80.99
204894_s_at	AOC3	−1.12	−14.71	1.0344e-35	80.92
225894_at	SYNPO2	−1.53	−14.69	1.2617e-35	80.72
210524_x_at	-NA-	−1.03	−14.69	1.2857e-35	80.70
227522_at	CMBL	−1.20	−14.63	2.3036e-35	80.12
221584_s_at	KCNMA1	−1.10	−14.62	2.6056e-35	80.00
219796_s_at	MUPCDH	−1.12	−14.60	3.2964e-35	79.76
220468_at	ARL14	−1.50	−14.59	3.4621e-35	79.72
205433_at	BCHE	−1.31	−14.54	5.5284e-35	79.25
215125_s_at	UGT1A6	−1.54	−14.52	7.0115e-35	79.01
208596_s_at	UGT1A3	−1.23	−14.49	9.1310e-35	78.75
227006_at	PPP1R14A	−1.32	−14.45	1.4384e-34	78.30
203060_s_at	PAPSS2	−1.18	−14.43	1.8090e-34	78.07
212592_at	IGJ	−1.90	−14.41	2.1252e-34	77.91
206199_at	CEACAM7	−2.58	−14.40	2.3506e-34	77.81
206576_s_at	CEACAM1	−1.43	−14.39	2.5103e-34	77.75
231975_s_at	MIER3	−1.07	−14.35	3.8508e-34	77.32
200795_at	SPARCL1	−1.29	−14.34	4.1568e-34	77.25
208791_at	CLU	−1.41	−14.25	1.0315e-33	76.34
209668_x_at	CES2	−1.08	−14.23	1.2269e-33	76.17
241994_at	XDH	−1.11	−14.19	1.9093e-33	75.73
211372_s_at	IL1R2	−1.09	−14.15	2.7350e-33	75.38
207126_x_at	UGT1A1	−1.11	−14.11	4.1654e-33	74.96
201957_at	PPP1R12B	−1.07	−14.07	6.1419e-33	74.57
225721_at	SYNPO2	−1.49	−14.04	8.5740e-33	74.24
205935_at	FOXF1	−1.01	−14.03	8.9357e-33	74.20
204532_x_at	UGT1A9	−1.04	−14.03	9.0571e-33	74.19
230595_at	LOC572558	−1.03	−13.97	1.6639e-32	73.58
226304_at	HSPB6	−1.06	−13.90	3.3027e-32	72.90
204326_x_at	MT1X	−1.10	−13.89	3.5362e-32	72.83
209283_at	CRYAB	−1.17	−13.84	5.7353e-32	72.35
203296_s_at	ATP1A2	−1.03	−13.76	1.3162e-31	71.53
204034_at	ETHE1	−1.02	−13.74	1.5735e-31	71.35
208383_s_at	PCK1	−1.85	−13.73	1.6976e-31	71.28
205267_at	POU2AF1	−1.30	−13.57	7.8874e-31	69.75
228232_s_at	VSIG2	−1.20	−13.56	9.0211e-31	69.62
224352_s_at	CFL2	−1.39	−13.56	9.0754e-31	69.61

Continued on Next Page...

Table D.6 – Continued

ProbeSetID	Symbol	Fold-Δ Log2	t statistic	P value (Bonf. corr.)	Likelihood
213921_at	SST	−1.09	−13.53	1.2534e-30	69.29
212097_at	CAV1	−1.23	−13.50	1.5862e-30	69.06
208450_at	LGALS2	−1.33	−13.48	1.8691e-30	68.90
203951_at	CNN1	−1.93	−13.45	2.5290e-30	68.60
212730_at	DMN	−1.96	−13.44	2.9249e-30	68.45
219508_at	GCNT3	−1.30	−13.39	4.8456e-30	67.95
224823_at	MYLK	−1.70	−13.36	6.4614e-30	67.67
204939_s_at	PLN	−1.68	−13.32	8.8539e-30	67.35
210302_s_at	MAB21L2	−1.90	−13.25	1.7306e-29	66.69
200621_at	CSRP1	−1.06	−13.25	1.7718e-29	66.66
226818_at	MPEG1	−1.02	−13.24	1.9208e-29	66.58
204508_s_at	CA12	−1.07	−13.12	6.0779e-29	65.44
219948_x_at	UGT2A3	−1.04	−13.11	6.9756e-29	65.30
217897_at	FXYD6	−1.00	−13.04	1.3500e-28	64.65
218087_s_at	SORBS1	−1.68	−13.00	2.0001e-28	64.26
228766_at	CD36	−1.30	−12.97	2.5291e-28	64.03
209114_at	TSPAN1	−1.06	−12.97	2.7166e-28	63.95
225728_at	SORBS2	−1.21	−12.92	4.2849e-28	63.50
203963_at	CA12	−1.24	−12.89	5.5175e-28	63.25
243278_at	FOXP2	−1.00	−12.87	7.0268e-28	63.01
203881_s_at	DMD	−1.13	−12.85	8.1353e-28	62.87
221748_s_at	TNS1	−1.57	−12.81	1.1958e-27	62.48
202388_at	RGS2	−1.17	−12.80	1.2623e-27	62.43
220645_at	FAM55D	−1.19	−12.79	1.4320e-27	62.31
208792_s_at	CLU	−1.13	−12.78	1.5876e-27	62.20
221747_at	TNS1	−1.23	−12.70	3.4615e-27	61.43
228202_at	PLN	−1.34	−12.69	3.6174e-27	61.39
202888_s_at	ANPEP	−1.68	−12.65	5.4081e-27	60.99
209948_at	KCNMB1	−1.10	−12.60	8.5174e-27	60.54
204897_at	PTGER4	−1.07	−12.60	8.7177e-27	60.51
224663_s_at	CFL2	−1.15	−12.57	1.1712e-26	60.22
213317_at	CLIC5	−1.04	−12.55	1.4504e-26	60.01
204940_at	PLN	−1.21	−12.48	2.6291e-26	59.42
202274_at	ACTG2	−1.94	−12.47	2.8400e-26	59.34
212192_at	KCTD12	−1.17	−12.45	3.4297e-26	59.16
210735_s_at	CA12	−1.04	−12.43	4.1354e-26	58.97
209498_at	CEACAM1	−1.30	−12.40	5.7253e-26	58.65
206664_at	SI	−1.55	−12.35	8.8947e-26	58.21
221667_s_at	HSPB8	−1.25	−12.35	9.0795e-26	58.19
220075_s_at	MUPCDH	−1.15	−12.32	1.2351e-25	57.88
202768_at	FOSB	−1.19	−12.27	1.9001e-25	57.46
211889_x_at	CEACAM1	−1.09	−12.22	2.8908e-25	57.04
217235_x_at	RPL14	−1.15	−12.15	5.9444e-25	56.33
217110_s_at	MUC4	−1.16	−12.09	9.9449e-25	55.82
214164_x_at	CA12	−1.07	−12.07	1.1945e-24	55.63
201324_at	EMP1	−1.06	−11.92	4.8947e-24	54.24
227727_at	MRGPRF	−1.03	−11.88	6.7304e-24	53.92
217148_x_at	IGL@	−1.25	−11.85	8.9939e-24	53.63
203240_at	FCGBP	−1.63	−11.84	9.5328e-24	53.58
217258_x_at	IVD	−1.12	−11.83	1.1377e-23	53.40
242601_at	LOC253012	−1.63	−11.81	1.3631e-23	53.22
228640_at	PCDH7	−1.15	−11.79	1.6055e-23	53.06
216984_x_at	RPL14	−1.14	−11.78	1.7425e-23	52.98
228133_s_at	NDE1	−1.69	−11.75	2.2827e-23	52.71
214598_at	CLDN8	−1.53	−11.75	2.3207e-23	52.69
238751_at	SORBS2	−1.05	−11.66	4.9471e-23	51.94

Continued on Next Page...

Table D.6 – Continued

ProbeSetID	Symbol	Fold-Δ Log2	t statistic	P value (Bonf. corr.)	Likelihood
215867_x_at	CA12	−1.09	−11.64	6.3560e-23	51.69
202555_s_at	MYLK	−1.50	−11.63	6.6548e-23	51.65
204895_x_at	MUC4	−1.14	−11.63	6.9011e-23	51.61
217179_x_at	LOC96610	−1.25	−11.61	7.9508e-23	51.47
217109_at	MUC4	−1.29	−11.61	8.1181e-23	51.45
205892_s_at	FABP1	−1.93	−11.51	1.9338e-22	50.59
212224_at	ALDH1A1	−1.18	−11.50	2.2125e-22	50.46
207245_at	UGT2B17	−2.42	−11.48	2.4808e-22	50.35
224342_x_at	LOC96610	−1.02	−11.43	3.9099e-22	49.89
217022_s_at	IGHA1	−1.62	−11.42	4.2874e-22	49.80
201058_s_at	MYL9	−1.60	−11.38	6.4665e-22	49.40
214433_s_at	SELENBP1	−1.17	−11.37	6.8334e-22	49.34
223484_at	C15orf48	−1.39	−11.35	7.9998e-22	49.18
201495_x_at	MYH11	−1.21	−11.34	8.9664e-22	49.07
202222_s_at	DES	−1.93	−11.28	1.5455e-21	48.53
214768_x_at	HLA-C	−1.42	−11.24	2.1947e-21	48.19
213953_at	KRT20	−1.55	−11.21	2.9274e-21	47.90
209374_s_at	IGHM	−1.39	−11.07	9.4651e-21	46.74
205097_at	SLC26A2	−2.11	−10.87	5.8246e-20	44.94
204938_s_at	PLN	−1.19	−10.81	9.7725e-20	44.43
224989_at	-NA-	−1.11	−10.77	1.3511e-19	44.10
214414_x_at	HBA1	−1.28	−10.75	1.6186e-19	43.93
209656_s_at	TMEM47	−1.04	−10.73	1.8414e-19	43.80
225782_at	MSRB3	−1.25	−10.72	1.9945e-19	43.72
227735_s_at	C10orf99	−1.37	−10.67	3.2477e-19	43.24
211645_x_at	-NA-	−1.35	−10.66	3.4472e-19	43.18
217378_x_at	-NA-	−1.10	−10.66	3.4824e-19	43.17
214027_x_at	DES	−1.14	−10.61	5.1641e-19	42.78
227736_at	C10orf99	−1.43	−10.52	1.1648e-18	41.97
211798_x_at	IGLJ3	−1.05	−10.50	1.4145e-18	41.78
211643_x_at	HLA-C	−1.09	−10.47	1.8470e-18	41.52
214777_at	-NA-	−1.21	−10.45	2.0943e-18	41.39
203980_at	FABP4	−1.30	−10.43	2.4336e-18	41.24
216207_x_at	IGKV1D-13	−1.03	−10.43	2.5820e-18	41.18
216576_x_at	NTN2L	−1.23	−10.39	3.5241e-18	40.88
211696_x_at	HBB	−1.17	−10.36	4.4393e-18	40.65
209116_x_at	HBB	−1.32	−10.29	8.2049e-18	40.04
206000_at	MEP1A	−1.16	−10.27	9.5067e-18	39.90
223597_at	ITLN1	−1.79	−10.27	9.6031e-18	39.89
216401_x_at	-NA-	−1.13	−10.26	1.0697e-17	39.78
202995_s_at	FBLN1	−1.07	−10.25	1.1281e-17	39.73
211644_x_at	HLA-C	−1.28	−10.25	1.1626e-17	39.70
207390_s_at	SMTN	−1.01	−10.14	2.8887e-17	38.80
207392_x_at	UGT2B15	−1.11	−10.11	3.6618e-17	38.56
214916_x_at	IL8	−1.10	−10.07	4.9357e-17	38.27
209210_s_at	FERMT2	−1.03	−10.04	6.5377e-17	37.99
211745_x_at	HBA2	−1.14	−9.91	1.9150e-16	36.93
217414_x_at	HBA1	−1.10	−9.78	5.5283e-16	35.88
210107_at	CLCA1	−1.71	−9.76	6.6334e-16	35.70
234764_x_at	IGLV1-44	−1.18	−9.70	1.0407e-15	35.25
205547_s_at	TAGLN	−1.31	−9.63	1.9156e-15	34.65
215176_x_at	NTN2L	−1.19	−9.61	2.2119e-15	34.51
217232_x_at	HBB	−1.06	−9.60	2.3203e-15	34.46
216510_x_at	ZCWPW2	−1.29	−9.59	2.6994e-15	34.31
209458_x_at	HBA1	−1.08	−9.57	3.1647e-15	34.15
213746_s_at	FLNA	−1.02	−9.55	3.7323e-15	33.99

Continued on Next Page...

Table D.6 – Continued

ProbeSetID	Symbol	Fold-Δ Log2	t statistic	P value (Bonf. corr.)	Likelihood
204018_x_at	HBA1	−1.03	−9.52	4.6754e-15	33.77
211699_x_at	HBA1	−1.05	−9.52	4.6964e-15	33.76
226302_at	ATP8B1	−1.03	−9.47	6.8365e-15	33.39
202291_s_at	MGP	−1.10	−9.28	3.0132e-14	31.93
225458_at	EXOC3	−1.31	−9.08	1.4833e-13	30.35
211896_s_at	DCN	−1.02	−9.00	2.7720e-13	29.74
204607_at	HMGCS2	−1.37	−8.99	3.0493e-13	29.64
204083_s_at	TPM2	−1.28	−8.92	5.0007e-13	29.15
211637_x_at	LOC652128	−1.08	−8.79	1.3340e-12	28.19
226654_at	MUC12	−1.16	−8.69	2.8573e-12	27.44
229659_s_at	PIGR	−1.15	−8.61	5.3622e-12	26.81
216491_x_at	IGHM	−1.11	−8.45	1.7640e-11	25.64
227725_at	ST6GALNAC1	−1.06	−8.10	2.2738e-10	23.12

D.4.4 Probesets upregulated in adenomas vs. cancer tissues

Table D.7: Probesets with increased expression in adenoma tissues relative to cancer tissues.

ProbeSetID	Symbol	Fold-Δ Log2	t statistic	P value (Bonf. corr.)	Likelihood
213106_at	ATP8A1	−1.32	−9.89	3.0825e-14	32.14
204811_s_at	CACNA2D2	−1.09	−9.71	1.0000e-13	31.01
228232_s_at	VSIG2	−1.60	−9.65	1.4994e-13	30.62
235976_at	SLITRK6	−1.37	−8.45	3.2762e-10	23.21
232481_s_at	SLITRK6	−1.67	−8.03	4.3327e-09	20.73
208063_s_at	CAPN9	−1.04	−7.93	7.7912e-09	20.16
214234_s_at	CYP3A5P2	−1.04	−7.64	4.6538e-08	18.44
223970_at	RETNLB	−2.05	−7.63	4.7545e-08	18.42
218211_s_at	MLPH	−1.16	−7.46	1.3608e-07	17.41
232176_at	SLITRK6	−1.33	−7.45	1.3748e-07	17.40
204508_s_at	CA12	−1.02	−7.39	2.0392e-07	17.02
205765_at	CYP3A5	−1.42	−7.33	2.8178e-07	16.71
214235_at	CYP3A5P2	−1.07	−7.33	2.9079e-07	16.68
223969_s_at	RETNLB	−1.84	−7.28	3.8401e-07	16.41
237521_x_at	-NA-	−1.07	−7.27	3.9314e-07	16.39
205259_at	NR3C2	−1.03	−7.19	6.3752e-07	15.93
215125_s_at	UGT1A6	−1.28	−6.81	5.5498e-06	13.85
236894_at	L1TD1	−1.30	−6.70	9.9335e-06	13.29
203963_at	CA12	−1.18	−6.69	1.0537e-05	13.24
204897_at	PTGER4	−1.18	−6.57	2.0624e-05	12.59
221874_at	KIAA1324	−1.03	−6.48	3.4133e-05	12.11
204607_at	HMGCS2	−1.95	−6.39	5.6210e-05	11.63
219543_at	PBLD	−1.03	−6.33	7.8191e-05	11.32
227719_at	-NA-	−1.19	−6.30	8.7743e-05	11.21
200884_at	CKB	−1.37	−6.23	0.0001	10.82
205927_s_at	CTSE	−1.40	−6.13	0.0002	10.33
208937_s_at	ID1	−1.44	−6.02	0.0003	9.79
203240_at	FCGBP	−1.97	−6.02	0.0004	9.75
210107_at	CLCA1	−2.42	−5.98	0.0004	9.60
215867_x_at	CA12	−1.01	−5.85	0.0009	8.94

Continued on Next Page...

Table D.7 – Continued

ProbeSetID	Symbol	Fold-Δ Log2	t statistic	P value (Bonf. corr.)	Likelihood
219955_at	L1TD1	−1.71	−5.78	0.0013	8.59
217110_s_at	MUC4	−1.08	−5.67	0.0022	8.09
231832_at	GALNT4	−1.02	−5.66	0.0024	8.02
226248_s_at	KIAA1324	−1.15	−5.59	0.0034	7.71
229070_at	C6orf105	−1.38	−5.58	0.0036	7.64
226302_at	ATP8B1	−1.14	−5.45	0.0070	7.02
227725_at	ST6GALNAC1	−1.59	−5.43	0.0077	6.94
242601_at	LOC253012	−1.60	−5.42	0.0079	6.91
214433_s_at	SELENBP1	−1.22	−5.41	0.0083	6.86
221841_s_at	KLF4	−1.07	−5.39	0.0090	6.79
204895_x_at	MUC4	−1.11	−5.24	0.0186	6.10
217109_at	MUC4	−1.30	−5.24	0.0191	6.07
227676_at	FAM3D	−1.00	−5.19	0.0242	5.85

D.4.5 Probesets upregulated in cancer vs. adenoma tissues

Table D.8: Probesets with increased expression in cancer tissues relative to adenoma tissues.

ProbeSetID	Symbol	Fold-Δ Log2	t statistic	P value (Bonf. corr.)	Likelihood
202404_s_at	COL1A2	3.26	14.42	9.5662e-28	62.03
202310_s_at	COL1A1	3.06	13.87	4.3976e-26	58.37
200665_s_at	SPARC	2.26	12.86	4.8381e-23	51.65
215076_s_at	COL3A1	2.40	12.41	1.0960e-21	48.65
202403_s_at	COL1A2	2.38	12.41	1.1059e-21	48.64
210495_x_at	FN1	2.82	12.39	1.3237e-21	48.47
212464_s_at	FN1	2.95	12.35	1.7306e-21	48.21
211719_x_at	FN1	2.97	12.33	2.0176e-21	48.07
216442_x_at	FN1	2.77	12.09	1.0254e-20	46.50
201852_x_at	COL3A1	2.40	11.86	5.1245e-20	44.96
211980_at	COL4A1	1.54	11.55	4.2921e-19	42.91
211161_s_at	COL3A1	2.38	11.04	1.4116e-17	39.55
225681_at	CTHRC1	3.01	10.98	2.0660e-17	39.18
201438_at	COL6A3	1.98	10.96	2.2951e-17	39.08
221729_at	COL5A2	2.26	10.67	1.6497e-16	37.18
212354_at	SULF1	2.24	10.58	3.1092e-16	36.57
210809_s_at	POSTN	2.84	10.54	4.0451e-16	36.32
221731_x_at	VCAN	2.17	10.36	1.3271e-15	35.17
211981_at	COL4A1	1.54	10.25	2.7382e-15	34.48
211964_at	COL4A2	1.48	10.22	3.3424e-15	34.28
218638_s_at	SPON2	1.21	10.00	1.4490e-14	32.87
202998_s_at	LOXL2	1.38	9.96	1.9842e-14	32.57
201744_s_at	LUM	2.22	9.67	1.3395e-13	30.73
201162_at	IGFBP7	1.34	9.57	2.5445e-13	30.11
204620_s_at	VCAN	1.98	9.50	3.9051e-13	29.70
227140_at	-NA-	2.33	9.49	4.2921e-13	29.61
201105_at	LGALS1	1.51	9.40	7.5051e-13	29.07
211959_at	IGFBP5	2.03	9.28	1.6322e-12	28.32
208788_at	ELOVL5	1.50	9.23	2.3595e-12	27.96
212667_at	SPARC	1.81	9.13	4.5101e-12	27.34

Continued on Next Page...

Table D.8 – Continued

ProbeSetID	Symbol	Fold-Δ Log2	t statistic	P value (Bonf. corr.)	Likelihood
221011_s_at	LBH	1.16	9.10	5.2477e-12	27.19
208782_at	FSTL1	1.33	9.07	6.5426e-12	26.98
213905_x_at	BGN	1.62	9.03	8.2377e-12	26.76
212489_at	COL5A1	1.60	8.99	1.0828e-11	26.50
217764_s_at	RAB31	1.56	8.90	1.8582e-11	25.98
225664_at	COL12A1	1.93	8.90	1.8642e-11	25.97
218468_s_at	GREM1	2.20	8.81	3.3449e-11	25.41
221730_at	COL5A2	1.92	8.79	3.7965e-11	25.29
217762_s_at	RAB31	1.55	8.79	3.8044e-11	25.29
212488_at	COL5A1	1.60	8.79	3.9604e-11	25.25
212353_at	SULF1	1.99	8.77	4.4131e-11	25.14
202311_s_at	COL1A1	2.19	8.71	6.3736e-11	24.79
226311_at	-NA-	1.40	8.70	6.8212e-11	24.72
210511_s_at	INHBA	1.70	8.70	6.8269e-11	24.72
203477_at	COL15A1	1.88	8.59	1.3776e-10	24.05
208851_s_at	THY1	1.02	8.56	1.6424e-10	23.88
207173_x_at	CDH11	1.90	8.52	2.1612e-10	23.61
217763_s_at	RAB31	1.54	8.49	2.5815e-10	23.44
213869_x_at	THY1	1.13	8.35	5.9205e-10	22.64
218469_at	GREM1	2.01	8.23	1.2859e-09	21.90
212344_at	SULF1	1.48	8.17	1.8063e-09	21.57
202450_s_at	CTSK	1.20	8.11	2.6688e-09	21.19
201069_at	MMP2	1.45	8.08	3.2316e-09	21.01
201185_at	HTRA1	1.29	8.02	4.7258e-09	20.64
211966_at	COL4A2	1.30	7.96	6.8467e-09	20.29
203083_at	THBS2	1.89	7.94	7.5303e-09	20.19
225799_at	LOC541471	1.03	7.93	8.1934e-09	20.11
226930_at	FNDC1	1.77	7.92	8.5662e-09	20.07
212077_at	CALD1	1.66	7.90	9.7001e-09	19.95
226237_at	COL8A1	1.94	7.84	1.3392e-08	19.64
201261_x_at	BGN	1.22	7.83	1.4369e-08	19.57
200832_s_at	SCD	1.26	7.80	1.7369e-08	19.39
231766_s_at	COL12A1	1.60	7.80	1.7732e-08	19.37
208850_s_at	THY1	1.13	7.79	1.8452e-08	19.33
209875_s_at	SPP1	2.68	7.77	2.0597e-08	19.23
224724_at	SULF2	1.22	7.72	2.8520e-08	18.91
201163_s_at	IGFBP7	1.30	7.72	2.9216e-08	18.89
224694_at	ANTXR1	1.64	7.54	8.4736e-08	17.87
231579_s_at	TIMP2	1.37	7.46	1.3621e-07	17.41
219087_at	ASPN	1.98	7.42	1.7142e-07	17.19
213428_s_at	COL6A1	1.21	7.38	2.0967e-07	17.00
200600_at	MSN	1.14	7.35	2.4529e-07	16.85
202878_s_at	CD93	1.00	7.31	3.2276e-07	16.58
203878_s_at	MMP11	1.05	7.30	3.4191e-07	16.53
205479_s_at	PLAU	1.04	7.29	3.4779e-07	16.51
201426_s_at	VIM	1.28	7.28	3.7003e-07	16.45
214247_s_at	DKK3	1.18	7.27	3.9891e-07	16.38
210095_s_at	IGFBP3	1.14	7.20	6.0092e-07	15.98
203325_s_at	COL5A1	1.11	7.18	6.6304e-07	15.89
209156_s_at	COL6A2	1.68	7.17	7.0655e-07	15.83
224560_at	TIMP2	1.27	7.15	8.2224e-07	15.68
209218_at	SQLE	1.15	7.08	1.1947e-06	15.32
202766_s_at	FBN1	1.34	7.03	1.5740e-06	15.06
201141_at	GPNMB	1.63	7.02	1.6790e-06	15.00
207191_s_at	ISLR	1.12	6.98	2.1562e-06	14.76
202859_x_at	IL8	2.04	6.98	2.1629e-06	14.76

Continued on Next Page...

Table D.8 – Continued

ProbeSetID	Symbol	Fold-Δ Log2	t statistic	P value (Bonf. corr.)	Likelihood
202237_at	NNMT	1.45	6.97	2.2642e-06	14.71
209955_s_at	FAP	1.28	6.95	2.5460e-06	14.60
211896_s_at	DCN	1.84	6.94	2.6224e-06	14.57
213125_at	OLFML2B	1.10	6.88	3.7044e-06	14.24
227566_at	HNT	1.11	6.87	4.0465e-06	14.15
201147_s_at	TIMP3	1.26	6.84	4.6120e-06	14.03
201150_s_at	TIMP3	1.33	6.83	4.8656e-06	13.98
204475_at	MMP1	2.33	6.83	4.9693e-06	13.96
233555_s_at	SULF2	1.04	6.82	5.2087e-06	13.91
208747_s_at	C1S	1.18	6.74	8.0206e-06	13.50
201792_at	AEBP1	1.26	6.70	1.0308e-05	13.26
204051_s_at	SFRP4	1.73	6.60	1.7526e-05	12.75
229802_at	-NA-	1.30	6.54	2.4214e-05	12.44
209395_at	CHI3L1	1.03	6.54	2.4476e-05	12.43
201893_x_at	DCN	1.37	6.51	2.9025e-05	12.27
209396_s_at	CHI3L1	1.15	6.42	4.8077e-05	11.78
215646_s_at	VCAN	1.43	6.40	5.1574e-05	11.72
201616_s_at	CALD1	1.36	6.40	5.3592e-05	11.68
37892_at	COL11A1	1.79	6.39	5.5501e-05	11.65
202238_s_at	NNMT	1.23	6.30	8.9097e-05	11.19
226694_at	AKAP2	1.10	6.28	9.9928e-05	11.08
201289_at	CYR61	1.28	6.26	0.0001	10.99
231879_at	COL12A1	1.34	6.25	0.0001	10.93
229218_at	COL1A2	1.06	6.24	0.0001	10.89
209596_at	MXRA5	1.19	6.12	0.0002	10.26
200974_at	ACTA2	1.07	6.09	0.0002	10.11
226777_at	-NA-	1.11	6.08	0.0002	10.09
211571_s_at	VCAN	1.30	6.08	0.0002	10.06
225710_at	GNB4	1.10	5.95	0.0005	9.41
209101_at	CTGF	1.16	5.93	0.0006	9.32
205547_s_at	TAGLN	1.59	5.92	0.0006	9.26
200986_at	SERPING1	1.07	5.87	0.0008	9.02
202283_at	SERPINF1	1.13	5.86	0.0009	8.97
204320_at	COL11A1	1.02	5.85	0.0009	8.94
217430_x_at	COL1A1	1.23	5.81	0.0011	8.73
218559_s_at	MAFB	1.04	5.79	0.0013	8.63
201667_at	GJA1	1.26	5.74	0.0016	8.40
232458_at	COL3A1	1.07	5.67	0.0023	8.07
223235_s_at	SMOC2	1.29	5.65	0.0025	7.99
203570_at	LOXL1	1.00	5.64	0.0026	7.94
211813_x_at	DCN	1.19	5.62	0.0029	7.86
204122_at	TYROBP	1.07	5.58	0.0035	7.66
223122_s_at	SFRP2	1.99	5.54	0.0043	7.48
201645_at	TNC	1.15	5.52	0.0049	7.36
234994_at	TMEM200A	1.10	5.50	0.0053	7.28
202620_s_at	PLOD2	1.06	5.41	0.0083	6.86
215049_x_at	CD163	1.04	5.39	0.0091	6.77
201859_at	SRGN	1.17	5.38	0.0094	6.74
210764_s_at	CYR61	1.06	5.32	0.0131	6.43
202917_s_at	S100A8	1.78	5.31	0.0134	6.41
203645_s_at	CD163	1.06	5.27	0.0166	6.21
201058_s_at	MYL9	1.38	5.26	0.0169	6.19
227099_s_at	LOC387763	1.03	5.21	0.0215	5.96
203382_s_at	APOE	1.14	5.20	0.0232	5.89
213524_s_at	G0S2	1.15	5.16	0.0271	5.74
201842_s_at	EFEMP1	1.14	5.14	0.0306	5.63

Continued on Next Page...

Table D.8 – Continued

ProbeSetID	**Symbol**	**Fold-Δ Log2**	t **statistic**	P **value (Bonf. corr.)**	**Likelihood**
204006_s_at	FCGR3B	1.04	5.10	0.0366	5.46
205828_at	MMP3	1.34	5.10	0.0369	5.45
202291_s_at	MGP	1.47	5.08	0.0397	5.38

Normal vs. Colitis

D.5 Hypothesis testing and validation

D.5.1 Validated differential display candidates

Table D.10: Validated biomarkers for neoplaisa discovered by differential display.

SeqID	*D* Value	Symbol	Fold-Δ	Sens-Spec
302	3.74	S100A11:LOC730558:LOC730278: ... more	3.86	0.97
66	3.15	SLC12A2	3.1	0.94
309	2.95	SLC12A2	3.19	0.93
296	2.79	APEX1	1.84	0.92
9	2.75	LOC731404:LOC729194:MYC	2.76	0.92
62	2.75	S100P	6.9	0.92
336	2.74	-NA-	2.98	0.91
20	2.69	-NA-	3.4	0.91
119	2.64	CCDC130:C19orf53	2.15	0.91
102	2.63	GALNT6:ELA1	3.36	0.91
263	2.63	NA:CG_63_Seq_ID263_st	1.99	0.91
56	2.56	-NA-	4.76	0.9
316	2.48	KIAA1199	3.88	0.89
110	2.47	SLC7A1:215979_s_at	1.98	0.89
7	2.39	KIAA1199	4.39	0.88
25	2.39	SLC7A1:215979_s_at	1.95	0.88
170	2.38	AOF2	1.67	0.88
234	2.32	GNB2L1:LOC647756	1.57	0.88
64	2.24	ETS2	2.07	0.87
80	2.19	LOC347509:LOC646641:LOC642451: ... more	1.7	0.86
4	2.17	TALDO1:C20orf199	2.93	0.86
280	2.15	LOC643412:BTF3	1.47	0.86
326	2.14	-NA-	1.66	0.86
186	2.11	DALRD3	1.52	0.85
239	2.09	OLFM2	1.4	0.85
192	2.08	GTPBP9	1.84	0.85
94	2.05	TMTC4:ERGIC3	1.62	0.85
195	2.05	DPEP1	1.74	0.85
255	2.05	RPL6:LOC343495:LOC139452: ... more	1.43	0.85
72	2.03	IFITM1	2.93	0.84
87	2.03	-NA-	1.56	0.84
271	2.03	PPM1G	1.56	0.84
304	2.02	ITGA6	2.01	0.84
233	2.01	IFITM2:IFITM3	2.53	0.84
256	1.99	NA:CG_85_Seq_ID256_st	1.34	0.84
318	1.98	-NA-	1.61	0.84
69	1.97	LOC649821:RPL14:RPL14L	1.38	0.84
52	1.95	MAGE:rs2072072_at	1.54	0.84
211	1.95	TGFBI	2.41	0.84
103	1.91	KIAA1199	2.17	0.83
154	1.91	TSPAN2	4.88	0.83
189	1.87	LOC730043:KRTCAP2	1.55	0.83
286	1.87	EIF3S2:DCDC2B:LOC648442: ... more	1.81	0.83

Continued on Next Page...

Table D.10 – Continued

SeqID	*D* Value	Symbol	Fold-Δ	Sens-Spec
5	1.86	TALDO1:C20orf199	1.2	0.82
185	1.86	LOC347509:LOC646641:LOC642451: . . . more	1.81	0.82
95	1.83	TTC7B:LOC729096:LOC96610: . . . more	5.12	0.82
319	1.82	-NA-	1.8	0.82
314	1.8	ZNF263	1.33	0.82
38	1.78	REG4	7.54	0.81
248	1.78	CYP2S1	1.91	0.81
262	1.78	NA:CG_74_Seq_ID262_st	1.28	0.81
220	1.77	HDGF	1.48	0.81
17	1.75	CADPS	1.55	0.81
47	1.75	PMS2L3	1.63	0.81
68	1.75	RPS4X:LOC650710:LOC400064: . . . more	1.59	0.81
100	1.73	RNF43	2.61	0.81
283	1.73	KCNQ1	1.51	0.81
294	1.72	NUBP1:LOC731361	1.4	0.81
281	1.71	REG4	7.21	0.8
312	1.7	HMGB1:LOC730825:LOC645292: . . . more	1.56	0.8
226	1.67	B4GALT3	1.56	0.8
208	1.66	LARP4:LOC730751:LOC728257	1.52	0.8
31	1.65	NA:CG_87_Seq_ID31_st	1.9	0.8
36	1.6	-NA-	1.34	0.79
55	1.56	MLLT3	1.59	0.78
223	1.55	PLAGL2:LOC152845:LOC649746	1.85	0.78
2	1.53	GIF	1.63	0.78
120	1.52	FNTB	1.34	0.78
76	1.51	-NA-	1.41	0.77
30	1.48	LRRFIP2	2.77	0.77
60	1.48	OLFM4	8.52	0.77
131	1.48	-NA-	1.53	0.77
285	1.46	-NA-	1.63	0.77
14	1.45	TTC7B:LOC729096:LOC96610: . . . more	4.06	0.77
51	1.45	GPR56	1.41	0.77
288	1.44	-NA-	1.56	0.76
74	1.4	GNL3L	1.35	0.76
82	1.4	LASS6	1.36	0.76
325	1.4	S100A6:228923_at	1.48	0.76
225	1.38	PFDN5	1.43	0.75
293	1.38	TMEM39B	1.31	0.75
287	1.37	ENO1	1.59	0.75
210	1.36	HN1L	1.64	0.75
143	1.35	SLC39A10:238968_at	1.86	0.75
146	1.33	ARMCX6	1.54	0.75
257	1.33	HSP90AA1:HSP90AA2	1.43	0.75
317	1.33	TM9SF1:238948_at	1.33	0.75
46	1.32	NA:CG_77_ZNF800_st	1.33	0.75
169	1.32	C14orf119	1.44	0.75
301	1.31	POMP	1.59	0.74
41	1.3	-NA-	1.45	0.74
237	1.3	DYNLRB1	1.38	0.74
246	1.3	NQO1	2.02	0.74
222	1.28	SPINK4	4.45	0.74
320	1.28	TPT1:LOC731557	1.15	0.74
12	1.27	-NA-	1.18	0.74
49	1.23	DEFA6	3.7	0.73
180	1.18	LRSAM1	1.13	0.72
40	1.16	C7orf38:ZKSCAN1	1.39	0.72
299	1.15	PRDX1	1.44	0.72

Continued on Next Page. . .

Table D.10 – Continued

SeqID	*D* Value	Symbol	Fold-Δ	Sens-Spec
111	1.13	COASY:ACACA	1.32	0.71
229	1.11	TM7SF3	1.3	0.71
105	1.09	NA:CG_77_ZNF800_st	1.37	0.71
278	1.09	GNAS:FGB	1.18	0.71
73	1.08	RPESP	1.67	0.71
216	1.08	DHX29:1566046_at	1.12	0.71
242	1.07	WARS	1.68	0.7
243	1.07	SPP1	1.76	0.7
247	1.07	EGFR	1.28	0.7
125	1.02	PLCB4	1.34	0.69
212	1.02	NA:CG_82_Seq_ID212_s_st	1.15	0.69
313	0.99	RNF130	1.32	0.69
177	0.98	TEX261	1.21	0.69
332	0.97	VAT1	1.27	0.69
264	0.96	NA:CG_65_Seq_ID264_st	1.21	0.68
85	0.95	GLT8D1	1.27	0.68
106	0.95	CCNI:LOC643280:LOC731020	1.34	0.68
324	0.92	ZNF223	1.11	0.68
338	0.9	REG1A	7.74	0.67
140	0.89	UGCGL2	1.29	0.67
259	0.89	BPHL	1.25	0.67
269	0.86	LDHB	1.8	0.67
107	0.85	NA:CG_64_Seq_ID107_st	1.47	0.66
182	0.85	PLEKHA8:PLEKHA9	1.21	0.66
303	0.85	OSBPL8:228985_at	1.29	0.66
157	0.83	RPRM	1.35	0.66
147	0.82	-NA-	1.3	0.66
260	0.82	NCK2:LOC729030	1.18	0.66
121	0.81	STARD3NL	1.17	0.66
295	0.81	FAT	1.15	0.66
133	0.8	C1orf123:MAGOH	1.16	0.66
193	0.8	C3orf19:TMEM135	1.12	0.66
266	0.79	-NA-	1.17	0.65
23	0.78	NA:CG_88_Seq_ID23_st	1.23	0.65
79	0.78	CCDC123	1.19	0.65
289	0.78	GPRC5A	1.36	0.65
228	0.76	LOC647047:TCP1:LOC400013: ... more	1.32	0.65
116	0.73	-NA-	1.17	0.64
142	0.73	GNG4	1.21	0.64
122	0.72	ORC2L	1.09	0.64
136	0.71	TTC7B:LOC729096:LOC96610: ... more	1.15	0.64
241	0.7	-NA-	1.31	0.64
311	0.7	GPSM3:PBX2:NOTCH4: ... more	1.26	0.64
96	0.68	FLJ25770:243875_at	1.36	0.63
174	0.68	LRPPRC	1.24	0.63
221	0.68	NA:CG_91_RPS26L_st	1.29	0.63
272	0.67	RETNLB	1.9	0.63
114	0.65	ALPK1	1.08	0.63
148	0.65	WAPAL	1.1	0.63
249	0.65	NA:CG_67_Seq_ID249_st	1.09	0.63
65	0.64	APOF:STAT2	1.14	0.63
19	0.63	TMEM39B	1.11	0.62
24	0.63	-NA-	1.08	0.62
151	0.63	C10orf112:1569954_at	1.19	0.62
202	0.62	ROD1	1.18	0.62
77	0.6	JPH3:EPSTI1	1.39	0.62
176	0.6	EHF	1.28	0.62

Continued on Next Page...

Table D.10 – Continued

SeqID	*D* Value	Symbol	Fold-Δ	Sens-Spec
265	0.58	WDR61:221532_s_at	1.24	0.61
45	0.57	LOC731933:DMBT1:LOC651581	2.3	0.61
83	0.57	PROS1:LOC648124	1.07	0.61
306	0.57	KIAA1370	1.18	0.61
18	0.56	MPZL1	1.06	0.61
21	0.56	NA:CG_5_Seq_ID21_st	1.07	0.61
196	0.53	RORA:NARG2:G3BP2	1.15	0.6
224	0.53	CALR	1.17	0.6
84	0.52	C9orf5	1.19	0.6
236	0.52	WIPF2:RARA:ARFGEF2	1.13	0.6
137	0.51	UCK2	1.13	0.6
297	0.51	PDE4DIP:GMEB1	1.1	0.6
93	0.49	COL8A1	1.17	0.6
134	0.49	PLCG2:HSPD1:LOC644745: ... more	1.17	0.6
108	0.48	TERF2	1.1	0.59

D.5.2 Adenoma specific biomarkers from differential display

Table D.11: Validated biomarkers for adenomas discovered by differential display.

SeqID	*D* Value	Symbol	Fold-Δ	Sens-Spec
154	3.45	TSPAN2	7.88	0.96
302	3.39	S100A11:LOC730558:LOC730278: ... more	3.89	0.95
9	3.24	LOC731404:LOC729194:MYC	3.12	0.95
66	3.2	SLC12A2	3.41	0.95
309	3.11	SLC12A2	3.63	0.94
20	3.03	-NA-	3.85	0.94
102	2.97	GALNT6:ELA1	3.82	0.93
296	2.94	APEX1	1.9	0.93
263	2.85	NA:CG_63_Seq_ID263_st	1.88	0.92
280	2.79	LOC643412:BTF3	1.55	0.92
186	2.76	DALRD3	1.7	0.92
336	2.75	-NA-	3.28	0.92
234	2.72	GNB2L1:LOC647756	1.65	0.91
170	2.71	AOF2	1.74	0.91
56	2.59	-NA-	5.39	0.9
95	2.59	TTC7B:LOC729096:LOC96610: ... more	6.87	0.9
316	2.59	KIAA1199	4.07	0.9
7	2.53	KIAA1199	4.63	0.9
62	2.52	S100P	7	0.9
256	2.48	NA:CG_85_Seq_ID256_st	1.4	0.89
110	2.4	SLC7A1:215979_s_at	1.84	0.88
239	2.4	OLFM2	1.47	0.88
36	2.39	-NA-	1.48	0.88
192	2.38	GTPBP9	1.86	0.88
326	2.35	-NA-	1.77	0.88

Continued on Next Page...

Table D.11 – Continued

SeqID	D Value	Symbol	Fold-Δ	Sens-Spec
4	2.33	TALDO1:C20orf199	2.97	0.88
25	2.33	SLC7A1:215979_s_at	1.86	0.88
103	2.32	KIAA1199	2.23	0.88
211	2.3	TGFBI	2.9	0.87
119	2.28	CCDC130:C19orf53	2.07	0.87
64	2.24	ETS2	2.14	0.87
14	2.2	TTC7B:LOC729096:LOC96610: . . . more	5.8	0.86
30	2.19	AUTS2	3.25	0.86
52	2.19	MAGE:rs2072072_at	1.66	0.86
318	2.19	-NA-	1.68	0.86
100	2.15	RNF43	2.65	0.86
5	2.12	TALDO1:C20orf199	1.49	0.86
69	2.1	RPL14:LOC647077:LOC649821: . . . more	1.41	0.85
38	2.05	REG4	9.28	0.85
255	1.98	RPL6:LOC343495:LOC139452: . . . more	1.43	0.84
51	1.97	GPR56	1.57	0.84
281	1.97	REG4	9.03	0.84
294	1.96	NUBP1:LOC731361	1.41	0.84
72	1.93	IFITM1:201601_x_at	2.54	0.83
55	1.91	MLLT3	1.69	0.83
195	1.91	DPEP1	1.57	0.83
223	1.9	PLAGL2:LOC152845:LOC649746	1.79	0.83
314	1.89	ZNF263	1.27	0.83
2	1.87	GIF	1.83	0.83
304	1.86	ITGA6	2.05	0.82
17	1.84	CADPS	1.55	0.82
189	1.84	LOC730043:KRTCAP2	1.59	0.82
272	1.8	RETNLB:223970_at	4.04	0.82
80	1.78	LOC347509:LOC646641:LOC642451: . . . more	1.5	0.81
87	1.78	-NA-	1.45	0.81
248	1.73	CYP2S1	1.94	0.81
94	1.72	TMTC4:ERGIC3	1.44	0.81
131	1.72	-NA-	1.63	0.81
233	1.69	IFITM2:201315_x_at	2.07	0.8
49	1.68	DEFA6	5.59	0.8
271	1.68	PPM1G	1.44	0.8
262	1.67	NA:CG_74_Seq_ID262_st	1.25	0.8
325	1.67	S100A6:228923_at	1.58	0.8
220	1.66	HDGF	1.44	0.8
320	1.63	TPT1:LOC731557	1.19	0.79
246	1.62	NQO1	2.17	0.79
60	1.61	OLFM4	12.26	0.79
222	1.6	SPINK4	6.88	0.79
46	1.59	NA:CG_77_ZNF800_st	1.38	0.79
68	1.57	RPS4X:LOC650710:LOC400064: . . . more	1.52	0.78
283	1.55	KCNQ1	1.46	0.78
319	1.51	-NA-	1.74	0.77
312	1.49	HMGB1:LOC645490:LOC645292: . . . more	1.46	0.77
82	1.46	LASS6:235463_s_at	1.39	0.77
185	1.44	LOC347509:LOC646641:LOC642451: . . . more	1.54	0.76
286	1.44	EIF3S2:DCDC2B:LOC648442: . . . more	1.58	0.76
266	1.42	-NA-	1.32	0.76
76	1.41	-NA-	1.36	0.76
208	1.41	LARP4:LOC730751:LOC728257	1.46	0.76
225	1.4	PFDN5	1.47	0.76
120	1.39	FNTB	1.31	0.76
146	1.39	ARMCX6	1.52	0.76

Continued on Next Page. . .

Table D.11 – Continued

SeqID	*D* Value	Symbol	Fold-Δ	Sens-Spec
278	1.39	GNAS:FGB	1.18	0.76
41	1.38	-NA-	1.47	0.75
226	1.36	B4GALT3	1.46	0.75
169	1.35	C14orf119	1.51	0.75
313	1.35	RNF130	1.45	0.75
264	1.34	NA:CG_65_Seq_ID264_st	1.32	0.75
216	1.33	DHX29:212648_at	1.15	0.75
47	1.32	PMS2L3	1.4	0.75
40	1.31	ZKSCAN1:214670_at	1.43	0.74
306	1.31	KIAA1370	1.39	0.74
147	1.23	-NA-	1.43	0.73
157	1.22	RPRM	1.44	0.73
105	1.21	NA:CG_77_ZNF800_st	1.39	0.73
285	1.2	-NA-	1.45	0.73
125	1.17	PLCB4	1.38	0.72
293	1.16	TMEM39B:218770_s_at	1.26	0.72
288	1.14	-NA-	1.37	0.72
317	1.13	TM9SF1:209150_s_at	1.27	0.71
180	1.12	LRSAM1	1.08	0.71
237	1.12	DYNLRB1	1.24	0.71
133	1.09	C1orf123:203197_s_at	1.22	0.71
31	1.08	NA:CG_87_Seq_ID31_st	1.46	0.71
143	1.08	SLC39A10:238968_at	1.36	0.71
297	1.08	PDE4DIP:GMEB1	1.21	0.71
311	1.08	GPSM3:PBX2:NOTCH4: ... more	1.43	0.71
221	1.05	NA:CG_91_RPS26L_st	1.48	0.7
193	1.04	C3orf19:TMEM135	1.29	0.7
287	1.03	ENO1:217294_s_at	1.38	0.7
324	1.03	ZNF223	1.11	0.7
111	1.01	ACACA:212186_at	1.27	0.69
73	1	RPESP	1.52	0.69
289	0.95	GPRC5A:203108_at	1.49	0.68
74	0.93	GNL3L	1.18	0.68
136	0.92	TTC7B:LOC729096:LOC96610: ... more	1.19	0.68
224	0.91	CALR	1.3	0.68
257	0.9	HSP90AA1:HSP90AA2	1.27	0.67
106	0.89	CCNI:LOC643280:LOC731020	1.3	0.67
229	0.89	TM7SF3	1.23	0.67
242	0.87	WARS:200628_s_at	1.58	0.67
332	0.86	VAT1:208626_s_at	1.24	0.67
301	0.85	POMP:217769_s_at	1.33	0.66
23	0.84	NA:CG_88_Seq_ID23_st	1.27	0.66
85	0.84	GLT8D1	1.24	0.66
247	0.82	EGFR	1.2	0.66
12	0.81	-NA-	1.1	0.66
210	0.81	HN1L	1.3	0.66
259	0.79	BPHL	1.22	0.65
121	0.78	STARD3NL	1.15	0.65
174	0.75	LRPPRC	1.25	0.65
330	0.75	CLCA1	2.21	0.65
253	0.74	-NA-	1.04	0.64
202	0.72	ROD1	1.21	0.64
107	0.7	NA:CG_64_Seq_ID107_st	1.4	0.64
177	0.68	TEX261:212083_at	1.13	0.63
35	0.67	KIAA1411	1.1	0.63
338	0.67	REG1A	4.68	0.63
79	0.66	CCDC123	1.15	0.63

Continued on Next Page...

Table D.11 – Continued

SeqID	*D* **Value**	**Symbol**	**Fold-Δ**	**Sens-Spec**
134	0.66	PLCG2:HSPD1:LOC644745: ... more	1.22	0.63
176	0.65	EHF	1.33	0.63
308	0.65	CLCA1	1.83	0.63

D.5.3 Common genes validated by custom and commercial probesets

Table D.12: Sequence IDs discovered by diffential display that were validated using both custom and commercial probesets. Note that several Sequence IDs appear to correspond to the same gene locus.

id	symbol	fold-raw	fold-other	Raw Sens/Spec	Other Sens/Spec
4	TALDO1:C20orf199	2.93	2.01	86.1	87.2
7	KIAA1199	4.39	25.16	88.4	93.7
38	REG4	7.54	6.46	81.3	81.7
45	LOC731933:DMBT1:LOC651581	2.3	2.09	61.2	64.9
60	OLFM4	8.52	8.54	77	79
62	S100P	6.9	4.22	91.5	93
66	SLC12A2	3.1	2.63	94.2	94.3
72	IFITM1	2.93	3.19	84.5	85.2
100	RNF43	2.61	3.35	80.6	92.8
102	GALNT6:ELA1	3.36	2.66	90.6	91.3
103	KIAA1199	2.17	25.16	83	93.7
211	TGFBI	2.41	3.69	83.5	91.1
222	SPINK4	4.45	4.47	73.9	74.8
233	IFITM2:IFITM3	2.53	2.14	84.3	82.9
246	NQO1	2.02	2.08	74.2	81.2
281	REG4	7.21	6.46	80.4	81.7
302	S100A11:LOC730558: ... more	3.86	3.21	96.9	97.4
304	ITGA6	2.01	2.15	84.4	91
309	SLC12A2	3.19	2.63	93	94.3
316	KIAA1199	3.88	25.16	89.3	93.7
338	REG1A	7.74	3.46	67.4	69.2

Table D.9: Gene set enrichment results using GSA library applied particular to the normal vs. inflamed (colitis) specimens. Inspection of the upregulated gene sets suggests increased expression in immunologically related pathways.

Downregulated Sets			
Name	Score	P-value	FDR
Butanoate metabolism	-0.6071	0	0.0%
Citrate cycle (TCA cycle)	-1.6223	0	0.0%
Fatty acid biosynthesis	-2.059	0	0.0%
Synthesis and degradation of ketone bodies	-1.9317	0	0.0%
Propanoate metabolism	-0.9897	0.00167	4.5%
Oxidative phosphorylation	-1.4408	0.0025	4.5%
Pyruvate metabolism	-0.8309	0.0025	4.5%
Ubiquinone biosynthesis	-1.5113	0.0025	4.5%
Valine, leucine, and isoleucine degradation	-0.6251	0.003	5.3%
Caprolactam degradation	-1.0655	0.00417	6.0%
Reductive carboyxlate cycle (CO2 fixation)	-1.235	0.005833	7.0%
Benzoate degradation via hydroxylation	-1.3311	0.00583	7.0%
ATP synthesis	-0.9791	0.0083	8.57%
Pentose and glucuronate interconversions	-1.3735	0.0083	8.57%
Alkaloid biosynthesis I	-1.3735	0.01083	10.4%
Sulfur metabolism	-1.0292	0.01417	12.75%
Fatty acid metabolism	-0.5789	0.01583	13.41%
Porphyrin and chlorophyll metabolism	-0.5796	0.0175	14.0%
Nitrogen metabolism	-0.4736	0.02	15.16%
Glyoxylate and dicarboxylate metabolism	-0.6309	0.0267	19.2%
Upregulated Sets			
Name	Score	P-value	FDR
Hematopoietic cell lineage	0.6318	0.0025	16.0%
Cell adhesion molecules (CAMs)	0.6502	0.003	16.0%
Cytokine-cytokine receptor interaction	0.4588	0.005	16.0%
T cell receptor signaling pathway	0.4417	0.005	16.0%
Toll-like receptor signalling pathway	0.5241	0.0067	16.0%
B cell receptor signaling pathway	0.5785	0.0067	16.0%
ECM-receptor interaction	0.7615	0.0083	17.14%
Jak-STAT signaling pathway	0.3002	0.0108	19.5%

D.5.4 Validated microarray discovered genes

Table D.13: Gene symbols observed to be differentially expressed in the custom microarrays by comparing either adenoma vs. normal tissues or cancer vs. normal.

GDF15	WDR72	LOC63928	MGC13057	SMPDL3A
SOX9	TCN1	UGP2	CRYAB	PDK4
NEBL	REG3A	PCK1	PLN	LMOD1
PDZK1IP1	TDGF1	ADH1B	PDCD4	NCLN
SOX4	UBD	CEACAM1	SPARCL1	PPP1R14A
AXIN2	GUCA1B	CES2	MYLK	CNN1
SLC12A2	CA4	EPB41L3	MYH11	ACTG2
FLJ37644	CEACAM7	PRDX6	HSD11B2	SRPX
LCN2	MS4A12	GPX3	MAOA	MATN2
RNF43	GUCA2B	SGK	CLEC3B	IGHG1
S100P	CLCA4	STMN2	ADAMDEC1	KCNMB1
SORD	CA1	CXCL12	PYY	SDPR
CDH3	AQP8	GCNT3	TNS1	CFD
ANXA3	TP53INP2	PKIB	GBA3	PPP1R12B
ENC1	SLC26A3	SEMA6D	CHRDL1	FAM129A
ASCL2	CD177	CLIC5	TSC22D3	C6orf204
TGFBI	OSTbeta	LPAAT-THETA	FHL1	KCTD12
RPL22L1	HPGD	CHGA	ABCA8	XDH
CXCL3	SLC4A4	TSPAN7	SFRP1	DMN
CCL20	CLDN23	CDKN2B	MIER3	MT1M
FOXQ1	ABCG2	SCNN1B	LRRC19	EDIL3
TACSTD2	EDN3	TRPM6	ANPEP	MGC14376
MMP7	CA2	SCARA5	DES	SPINK5
KIAA1199	SDCBP2	AKR1B10	RPL24	HSPB8
LGR5	MGC4172	HSD17B2	ANGPTL1	FGL2
MET	DHRS9	TCF21	MYL9	CFL2
SLC6A6	MALL	EMP1	ADH1C	CAV1
SERPINB5	XLKD1	DPT	PRIMA1	MT2A
DPEP1	ZG16	ACAT1	GCG	CD36
TESC	SLC26A2	CITED2	SYNPO2	
MSLN	PLAC8	SEPP1	CLDN8	

D.5.5 Validated biomarkers discriminating adenoma vs. cancer

Table D.14: Microarray-discovered biomarkers upregulated in adenomas relative to cancer tissues that were likewise differentially expressed in validation data

ProbesetID	Symbol	Fold-$Delta$ (Log2)	t statistic	P Value (Bonf Corr)	Likelihood
210107_at	CLCA1	−3.32	−7.31	1.5862e-06	10.11
223970_at	RETNLB	−1.86	−5.60	0.0001	4.94
228232_s_at	VSIG2	−1.14	−5.03	0.0003	3.22
205765_at	CYP3A5	−0.85	−4.94	0.0003	2.94
203240_at	FCGBP	−1.62	−4.90	0.0003	2.83
223969_s_at	RETNLB	−1.22	−4.65	0.0006	2.07
242601_at	LOC253012	−1.11	−4.52	0.0008	1.69
226248_s_at	KIAA1324	−1.11	−4.27	0.0011	0.97
227676_at	FAM3D	−1.01	−4.16	0.0014	0.66
219955_at	L1TD1	−1.71	−4.04	0.0019	0.32
218211_s_at	MLPH	−0.67	−3.91	0.0024	−0.03

215867_x_at	CA12	−0.89	−3.80	0.0029	−0.34
227725_at	ST6GALNAC1	−0.65	−3.72	0.0033	−0.56
200884_at	CKB	−1.17	−3.69	0.0034	−0.64
204607_at	HMGCS2	−1.56	−3.63	0.0038	−0.80
232481_s_at	SLITRK6	−1.11	−3.61	0.0039	−0.86
204508_s_at	CA12	−0.85	−3.60	0.0039	−0.89
203963_at	CA12	−0.87	−3.55	0.0043	−1.01
214234_s_at	CYP3A5P2	−0.77	−3.51	0.0048	−1.13
214433_s_at	SELENBP1	−1.00	−3.13	0.0122	−2.10
231832_at	GALNT4	−0.35	−3.10	0.0128	−2.17
221841_s_at	KLF4	−0.39	−2.75	0.0253	−3.01
219543_at	PBLD	−0.42	−2.74	0.0253	−3.04
204897_at	PTGER4	−0.23	−2.64	0.0312	−3.27
208937_s_at	ID1	−0.73	−2.44	0.0442	−3.70

D.5.6 Validated biomarkers elevated in cancers relative to adenomas

Table D.15: Microarray-discovered biomarkers upregulated in cancers relative to adenoma tissues that were likewise differentially expressed in validation data

ProbesetID	Symbol	Fold-$Delta$ (Log2)	t statistic	P Value (Bonf Corr)	Likelihood
208850_s_at	THY1	0.74	5.32	0.0002	4.10
203878_s_at	MMP11	0.36	5.08	0.0003	3.36
225664_at	COL12A1	0.51	4.92	0.0003	2.87
217430_x_at	COL1A1	1.09	4.91	0.0003	2.86
226311_at	-NA-	0.47	4.90	0.0003	2.83
209396_s_at	CHI3L1	1.15	4.47	0.0008	1.57
202310_s_at	COL1A1	1.21	4.46	0.0008	1.54
212489_at	COL5A1	0.84	4.44	0.0008	1.47
211966_at	COL4A2	0.67	4.44	0.0008	1.47
231879_at	COL12A1	1.10	4.42	0.0008	1.40
208851_s_at	THY1	0.46	4.37	0.0009	1.26
213869_x_at	THY1	0.37	4.34	0.0009	1.18
207191_s_at	ISLR	0.47	4.24	0.0011	0.89
211981_at	COL4A1	0.70	4.04	0.0019	0.33
209395_at	CHI3L1	1.24	3.95	0.0023	0.06
231766_s_at	COL12A1	0.73	3.94	0.0023	0.05
211980_at	COL4A1	0.79	3.90	0.0025	−0.08
201645_at	TNC	0.44	3.84	0.0028	−0.24
203477_at	COL15A1	0.50	3.82	0.0029	−0.30
221731_x_at	VCAN	0.47	3.81	0.0029	−0.33
202311_s_at	COL1A1	0.90	3.77	0.0031	−0.43
205479_s_at	PLAU	0.45	3.74	0.0032	−0.50
203325_s_at	COL5A1	0.82	3.71	0.0033	−0.59
204620_s_at	VCAN	0.49	3.68	0.0034	−0.66
213905_x_at	BGN	0.40	3.67	0.0034	−0.69
221729_at	COL5A2	0.31	3.60	0.0039	−0.88
201261_x_at	BGN	0.30	3.48	0.0052	−1.22
202404_s_at	COL1A2	0.66	3.25	0.0095	−1.80
210495_x_at	FN1	0.45	3.23	0.0098	−1.85
208788_at	ELOVL5	0.22	3.19	0.0108	−1.96
212488_at	COL5A1	0.41	3.18	0.0109	−1.99
215646_s_at	VCAN	0.67	3.08	0.0135	−2.23
212344_at	SULF1	0.71	3.03	0.0152	−2.36
209955_s_at	FAP	0.27	2.98	0.0170	−2.48
211964_at	COL4A2	0.49	2.95	0.0179	−2.54

202238_s_at	NNMT	0.21	2.94	0.0179	−2.57
216442_x_at	FN1	0.48	2.94	0.0179	−2.57
221730_at	COL5A2	0.53	2.91	0.0189	−2.64
210511_s_at	INHBA	0.43	2.89	0.0194	−2.68
204051_s_at	SFRP4	0.20	2.88	0.0194	−2.70
211571_s_at	VCAN	0.46	2.88	0.0194	−2.71
219087_at	ASPN	0.46	2.85	0.0203	−2.77
227566_at	HNT	0.17	2.85	0.0203	−2.77
218638_s_at	SPON2	0.41	2.77	0.0248	−2.97
211719_x_at	FN1	0.46	2.75	0.0253	−3.02
202403_s_at	COL1A2	0.72	2.74	0.0253	−3.03
201792_at	AEBP1	0.24	2.70	0.0272	−3.12
200665_s_at	SPARC	0.48	2.67	0.0293	−3.20
210809_s_at	POSTN	0.57	2.62	0.0323	−3.31
233555_s_at	SULF2	0.42	2.57	0.0361	−3.42
212354_at	SULF1	0.56	2.56	0.0362	−3.44
201438_at	COL6A3	0.39	2.53	0.0386	−3.51
212353_at	SULF1	0.14	2.52	0.0389	−3.52
217764_s_at	RAB31	0.19	2.50	0.0405	−3.57
201289_at	CYR61	0.20	2.48	0.0416	−3.61
212464_s_at	FN1	0.49	2.45	0.0441	−3.67
202998_s_at	LOXL2	0.42	2.45	0.0441	−3.68
229218_at	COL1A2	0.16	2.44	0.0442	−3.69

D.5.7 Validation of turned-off biomarkers

Table D.16: Putative turned-OFF biomarkers that also showed decreased neoplastic expression in the validation data

ProbeSetID	**Symbol**	**Fold-Δ Log2**	***t* statistic**	***P* value (Bonf. corr.)**	**Likelihood**
211549_s_at	HPGD	-1.35	-11.47	4.5740e-16	29.56
228706_s_at	CLDN23	-1.66	-11.17	7.5963e-16	28.36
220037_s_at	XLKD1	-1.46	-10.15	3.0489e-14	24.28
220812_s_at	HHLA2	-1.77	-9.75	1.1984e-13	22.63
209613_s_at	ADH1B	-0.90	-8.54	1.4371e-11	17.64
235146_at	No Symbol	-1.06	-8.14	6.4608e-11	15.96
224412_s_at	TRPM6	-0.87	-7.57	5.9836e-10	13.60
207980_s_at	CITED2	-1.05	-7.44	9.0460e-10	13.06
207080_s_at	PYY	-1.38	-7.05	4.0812e-09	11.45
204931_at	TCF21	-0.47	-6.74	1.3578e-08	10.16
220376_at	LRRC19	-0.64	-5.74	7.1758e-07	6.15
238751_at	SORBS2	-1.07	-5.47	1.9180e-06	5.10
204719_at	ABCA8	-0.32	-5.35	2.7799e-06	4.66
228885_at	RPL24	-0.73	-5.03	8.8044e-06	3.46
214598_at	CLDN8	-0.44	-4.64	3.5821e-05	2.03
231773_at	ANGPTL1	-0.38	-4.06	0.0002	0.04
222717_at	SDPR	-0.17	-3.68	0.0009	-1.19
206637_at	P2RY14	-0.13	-3.14	0.0046	-2.78
228766_at	CD36	-0.14	-2.98	0.0069	-3.21
202920_at	ANK2	-0.21	-2.78	0.0115	-3.72
204940_at	PLN	-0.18	-2.65	0.0157	-4.05
231925_at	P2RY1	-0.10	-2.47	0.0240	-4.47
230788_at	GCNT2	-0.13	-2.45	0.0240	-4.51

D.5.8 ROC curves for novel genes

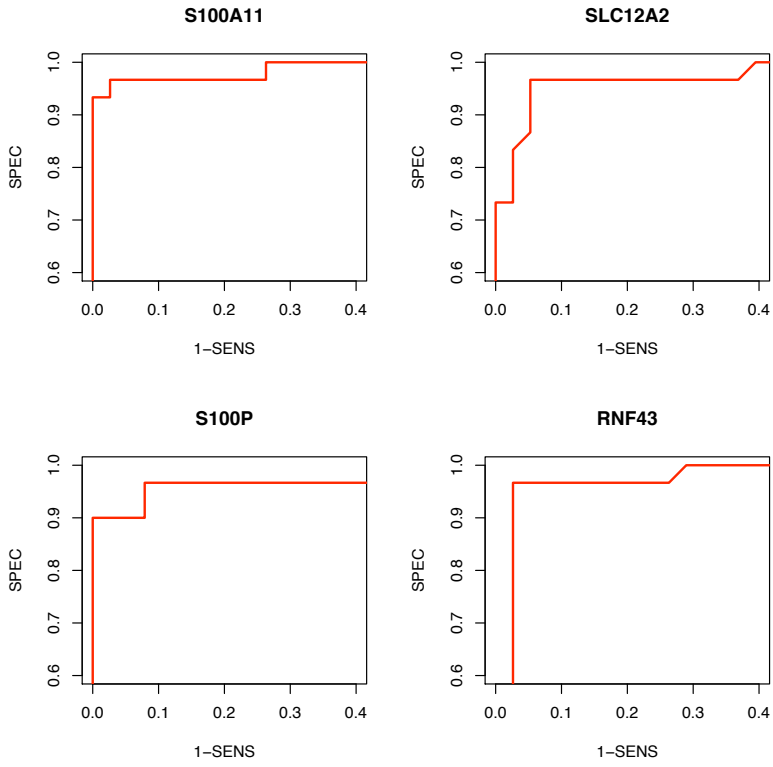


Figure 4.8: ROC curves for novel genes which were validated following consistent discovery in both the differential display research and the microarray experiments (figure 1 of 4).

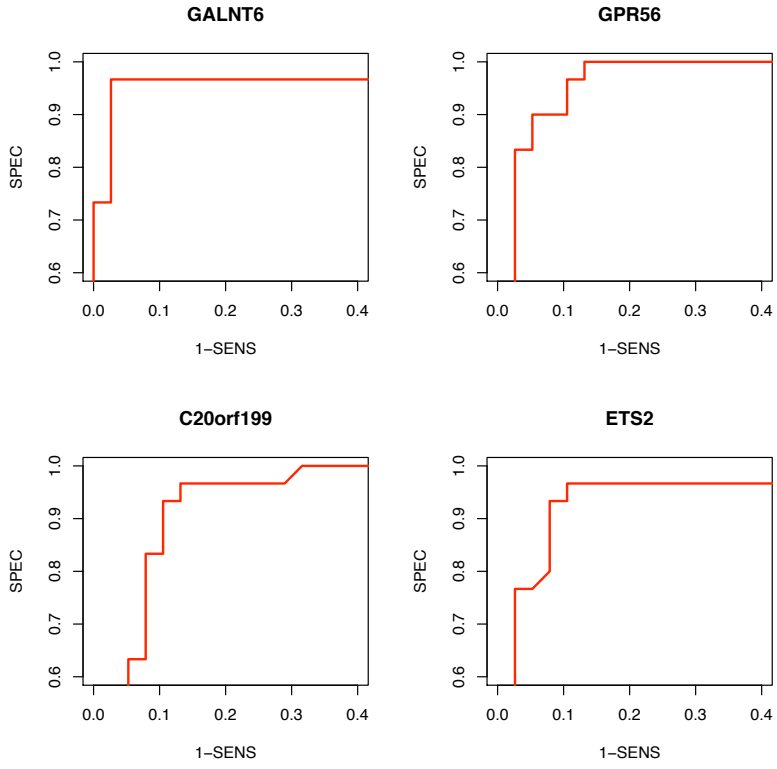


Figure 4.9: ROC curves for novel genes which were validated following consistent discovery in both the differential display research and the microarray experiments (figure 2 of 4).

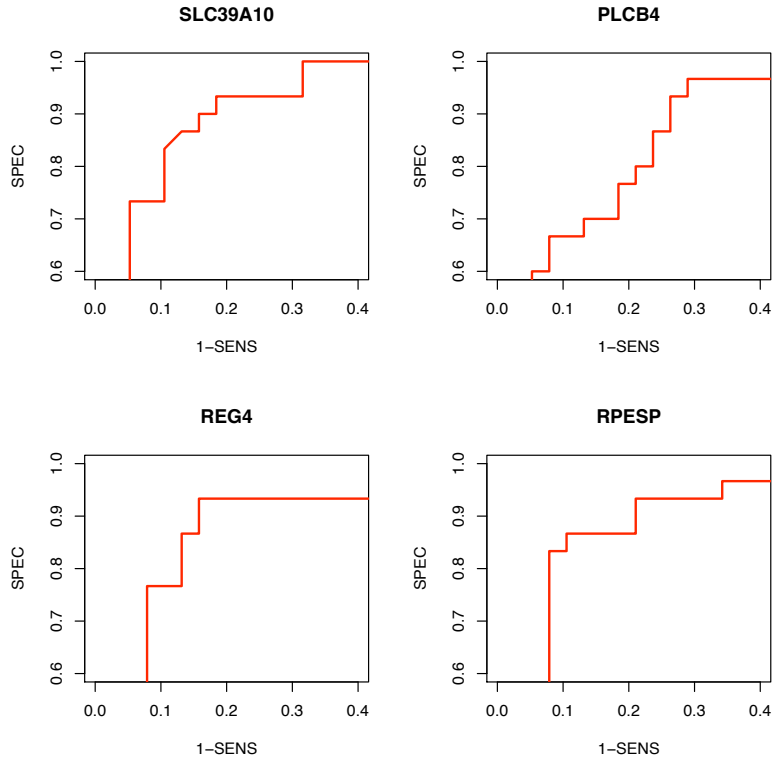


Figure 4.10: ROC curves for novel genes which were validated following consistent discovery in both the differential display research and the microarray experiments (figure 3 of 4).

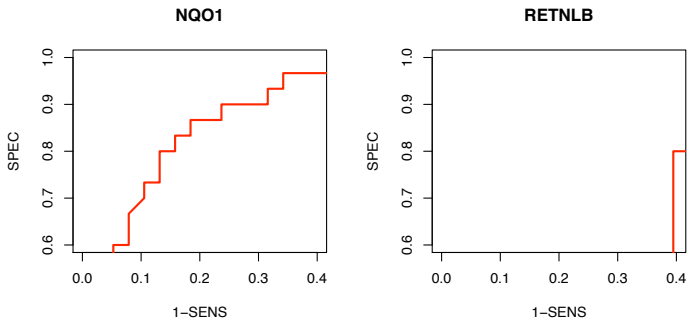


Figure 4.11: ROC curves for novel genes which were validated following consistent discovery in both the differential display research and the microarray experiments (figure 4 of 4).

D.5.9 List of validated genes

Table D.17: All validated genes from all sources of discovery: diff. display, microarray, and literature candidates

A2M	CCDC123	DES	GNL3L	LGR5	NEBL	RELL1	ST6GALNAC1
ABCA8	CCDC80	DHRS9	GPA33	LILRB1	NEXN	RETNLB	STMN2
ABCG2	CCL15	DHX29	GPM6B	LMOD1	NFE2L3	RFC3	SULF1
ABI3BP	CCL20	DIAPH2	GPNMB	LOC253012	NLF1	RHOQ	SULT1A1
ABP1	CCL28	DIP13B	GPR56	LOC285382	NPDC1	RNF12	SYNPO2
ACACA	CCL8	DKFZP564O0823	GPRC5A	LOC389634	NQO1	RNF130	TACSTD2
ACAT1	CCND1	DMN	GPSM3	LOC401022	NR3C1	RNF43	TAGLN
ACTA2	CCNI	DNASE1L3	GPX2	LOC541471	NR3C2	ROD1	TBC1D9
ACTG2	CD14	DPEP1	GPX3	LOC554203	NUBP1	RP11-50D16.3	TBX3
ADAMDEC1	CD163	DPT	GREM1	LOC63928	No Symbol	RP5-875H10.1	TCEA3
ADAMTS1	CD177	DSCR1	GSTM5	LOC646627	OGN	RPESP	TCF21
ADH1B	CD36	DUOX2	GTF3A	LOXL2	OGT	RPL13	TCN1
ADH1C	CD44	DUSP1	GTPBP9	LPAAT-THETA	OLFM4	RPL14	TDGF1
AKAP12	CD55	DUSP27	GUCA1B	LRPPRC	ORC2L	RPL22L1	TERF2
AKR1B10	CDA	DUSP5	GUCA2B	LRRC19	OSBPL8	RPL24	TESC
ALDH1A1	CDC2	DYNC1LI2	GUCY1A3	LRRFIP2	OSTbeta	RPL6	TEX261
ALPK1	CDCA7	DYNLRB1	H19	LRSAM1	P2RY14	RPL7L1	TEX9
ANGPTL1	CDH11	ECT2	HDGF	LSM3	PADI2	RPS4X	TGFBI
ANK2	CDH3	EDG2	HECTD1	LUM	PAG1	RPS7	TGIF
ANK3	CDKL1	EDIL3	HEPH	LY6G6D	PAICS	RRM2	TIMP1
ANLN	CDKN2B	EDN3	HHLA2	MAB21L2	PAPSS2	S100A11	TIMP2
ANPEP	CEACAM1	EFEMP1	HIG2	MAFB	PAQR5	S100A2	TIMP3
ANXA3	CEACAM7	EHF	HIGD1A	MAGEF1	PBLD	S100A6	TM7SF3

AOC3	CEL	EIF3S2	HMGB1	MALL	PBX3	S100P	TM9SF1	
AOF2	CES2	EMP1	HMGB3	MAML2	PCCA	SCARA5	TM9SF3	
AP1S2	CFD	ENAM	HMGCS2	MAOA	PCK1	SCARNA17	TMEM39B	
APEX1	CFL2	ENC1	HN1L	MAP3K5	PDCD4	SCD	TMEM97	
APOBEC1	CHGA	ENO1	HOXA9	MATN2	PDE6A	SCNN1B	TMEPAI	
APP	CHI3L1	ENTPD5	HOXB6	MEIS1	PDE9A	SCYL1	TNS1	
AQP8	CHRDL1	EPB41L3	HPGD	MEP1A	PDK4	SDC2	TP53INP2	
ARL14	CITED2	ERGIC3	HSD11B2	MET	PDLIM3	SDCBP2	TPSAB1	
ARMCX6	CKB	ERO1L	HSD17B2	METTL7A	PDZK1IP1	SDCCAG1	TPSB2	
ASAHL	CKS2	ETHE1	HSP90AA1	MFAP2	PECI	SDPR	TPX2	
ASCL2	CLCA1	ETS2	HSPA1A	MFAP4	PEX7	SELENBP1	TRAF5	
ASH1L	CLCA4	EXO1	HSPB6	MFSD4	PFDN4	SEMA6A	TRIM29	
ASPN	CLDN1	EXOC3	HSPB8	MGC13057	PFDN5	SEMA6D	TRPM6	
ATP10B	CLDN23	F13A1	HSPH1	MGC14376	PHACTR2	SEPHS1	TSC22D3	
ATP8B1	CLDN8	F2RL1	IDH3A	MGC4172	PHF14	SEPP1	TSPAN1	
AUTS2	CLEC3B	FABP1	IFITM1	MGP	PHLDA1	SERPINA1	TSPAN2	
AXIN2	CLIC5	FABP6	IFITM2	MIA3	PIGR	SERPINB5	TSPAN7	
AZGP1	CLU	FAM105A	IGFBP2	MIER3	PKIB	SERPINE2	TST	
B4GALT3	CMBL	FAM107A	IGHG1	MLH1	PLAC8	SF3B1	TTC28	
BACE2	CMIP	FAM129A	IL1R2	MLLT3	PLAGL2	SFRP1	TTRAP	
BCAS1	CNIH	FAM20B	IL8	MLPH	PLAU	SFRP2	TUBB6	
BCMP11	CNN1	FAM3D	INHBA	MMP1	PLCB4	SGCE	UBD	
BEST2	CNOT2	FAM55D	INSR	MMP11	PLCE1	SGK	UBE2C	
BGN	COL11A1	FAM84A	IQGAP2	MMP12	PLEKHA8	SI	UBE2S	
BLNK	COL12A1	FAP	ISX	MMP28	PLN	SLC11A2	UCK2	
BMS1L	COL1A1	FAT	ITGA6	MMP3	PLOD2	SLC12A2	UGCGL2	
BPHL	COL1A2	FBLN1	ITLN1	MMP7	PMS2L3	SLC16A9	UGDH	
BTF3	COL3A1	FCGBP	ITM2A	MPEG1	POF1B	SLC20A1	UGP2	

BTNL8	COL5A1	FGFR2	ITM2C	MPZL1	POLR1A	SLC26A2	UGT1A1
C10orf99	COL5A2	FGL2	JMJD1C	MRC1	POMP	SLC26A3	UGT1A3
C14orf119	COL6A1	FHL1	KCNMA1	MRGPRF	POSTN	SLC39A10	UGT1A6
C14orf58	COL6A2	FKBP5	KCNMB1	MRPS25	POU2AF1	SLC44A4	UGT1A8
C14orf94	COL6A3	FLJ21511	KCNN4	MS4A12	PPAP2A	SLC4A4	UGT1A9
C15orf48	COL8A1	FLJ25770	KCNQ1	MS4A4A	PPAP2B	SLC6A6	UGT2B15
C18orf1	COMP	FLJ37644	KCTD12	MSH3	PPM1G	SLC7A1	UGT2B17
C19orf53	CPA6	FLNA	KHDRBS1	MSLN	PPP1R12B	SLC7A5	UHRF2
C1S	CRYAB	FMO5	KIAA0460	MSTO1	PPP1R14A	SLCO1B3	VAMP3
C1orf115	CSE1L	FN1	KIAA0828	MT1F	PRDX1	SLCO4A1	VAT1
C1orf123	CSPG2	FNBP1	KIAA1199	MT1G	PRDX6	SLITRK6	VIM
C20orf118	CSRP1	FOXF1	KIAA1370	MT1H	PRIMA1	SMPDL3A	VSIG2
C20orf199	CST1	FOXF2	KIAA1411	MT1M	PRKACB	SNHG10	WAPAL
C20orf42	CTSE	FOXQ1	KIAA1600	MT1X	PRNP	SNTB2	WARS
C3orf19	CXCL1	FTH1	KIFAP3	MT2A	PSAT1	SOD1	WDR51B
C3orf28	CXCL12	FUCA1	KLF4	MTHFD1L	PTEN	SORBS1	WDR72
C4orf34	CXCL2	FXYD3	KLK10	MUC12	PTGER4	SORBS2	XDH
C6orf105	CXCL3	FXYD6	KLK11	MUC2	PTP4A1	SORD	XLKD1
C6orf204	CXCL5	G3BP2	KRT20	MUC4	PTRF	SOX4	ZG16
C9orf19	CYB5B	GALNT6	KRT23	MUCDHL	PUS7	SOX9	ZKSCAN1
C9orf5	CYBRD1	GBA3	KRTCAP2	MXD1	PVT1	SPARC	ZNF223
CA1	CYP2S1	GCG	L1TD1	MYC	PYY	SPARCL1	ZNF263
CA12	CYP3A5	GCNT2	LAMA1	MYH11	QPCT	SPINK1	ZNF447
CA2	CYR61	GCNT3	LARP4	MYL9	RAB27A	SPINK4	ZNRD1
CA4	DACH1	GDF15	LASS6	MYLK	RARRES2	SPINK5	ZNRF3
CADPS	DALRD3	GGT6	LCN2	MYO1A	RBMS1	SPON1	
CALD1	DCN	GLT8D1	LDHAL6B	MYO5B	RDHE2	SPP1	
CALM1	DDIT4	GMDS	LDHB	NCK2	REG1A	SQRDL	

CALR	DDR2	GMEB1	LEF1	NCLN	REG1B	SRI
CASP7	DEFA5	GNAS	LGALS2	NDE1	REG3A	SRPX
CAV1	DEFA6	GNB2L1	LGALS4	NDRG1	REG4	SST

Appendix E

Appendix: Publications and Patents Arising

E.1 Peer reviewed articles

Lawrence LaPointe, Robert Dunne, Glenn S Brown, Daniel L Worthley, Peter L. Molloy, David Wattchow, and Graeme P. Young. Map of differential transcript expression in the normal human large intestine. *Physiol. Genomics*, 33(1):50–64, 2008

E.2 Invited talks

1. Lawrence LaPointe. The normal colon gene map: from maths to genes. In *Australian Gastroenterology Week*, Adelaide, 2006. Invited Session Presentation
2. Lawrence LaPointe. Biomarkers for colorectal neoplasia. In *M D Anderson Cancer Center*, Houston, TX USA, 2007a. Invited Seminar
3. Lawrence LaPointe. Brave new world: Advances in genomics; gene expression mapping of the normal colon. In *New Zealand Bio*, Auckland, NZ, 2007b. Invited Session Presentation

E.3 Conference posters

1. Lawrence LaPointe, Robert Dunne, Peter Molloy, L Clark, Thu Ho, Susanne Pedersen, and Graeme P Young. Biomarkers with high sensitivity and specicity for colorectal adenomas and carcinomas. In *Gastroenterology*. AGA, 2009. Poster: DDW (Chicago)
2. Susanne Pedersen, Glenn Brown, Lloyd Graham, Robert Dunne, Peter Molloy, L Clark, Graeme P Young, and Lawrence LaPointe. A novel colorectal neoplasia gene (crng) with high sensitivity and specificity for both adenomas and cancers. In *Gastroenterology*. AGA, 2009a. Poster: DDW (Chicago)
3. Susanne Pedersen, Emma Richards, Aidan McEvoy, Robert Dunne, Glenn Brown, L Clark, Graeme P Young, and Lawrence LaPointe. Alternative splicing of s100a11 in colorectal adenomas and carcinomas. In *Gastroenterology*. AGA, 2009b. Poster: DDW (Chicago)
4. A Moynihan, P Molloy, V Papangelis, Graeme Young, and Lawrence LaPointe. Upregulation of mesothelin, regiv, and transcobalamin in colon adenomas and cancer. In *Gastroenterology*. AGA, 2009. Poster: DDW (Chicago)
5. Lawrence LaPointe and Robert Dunne. Normalization of custom microarrays. In *AMATA 2007 Meeting*, 2007. Poster: AMATA (Brisbane)
6. H Kiiveri, Robert Dunne, and Lawrence LaPointe. Canonical variate analysis and microarrays. In *AMATA 2005 Meeting*, 2005. Poster: AMATA (Adelaide)
7. Lawrence LaPointe and Robert Dunne. Comparison of machine learning techniques to identify biomarkers for colorectal cancer in publicly available data. In *International Society of Computational Biology*, 2005c. Poster: ISMB (Detroit, USA)
8. Lawrence LaPointe and Robert Dunne. Identification of colorectal cancer biomarkers using publicy available gene expression data. In *Gastroenterology*. AGA, 2005b. Poster: DDW (Chicago)
9. Lawrence LaPointe, Graeme P Young, and Howard Chandler. Analysis of mrna expression profiles in colorectal adenomas using k-nearest neighbor

cluster analysis. In *Gastroenterology*. AGA, 2005b. Poster: DDW (Orlando)

E.4 Patents submitted

1. R. James. Nucleic acid markers for use in determining predisposition to neoplasm and/or adenoma, 2001
2. Lawrence LaPointe and Robert Dunne. A method of diagnosis: markers of anatomical location, 2005d
3. Lawrence LaPointe, R Dunne, G Young, T Lockett, B Wilson, and P Molloy. Nucleic acid markers for use in determining predisposition to neoplasm and/or adenoma, 2007a
4. Lawrence LaPointe, Robert Dunne, Graeme Young, Peter Molloy, Trevor Lockett, and William Wilson. A method of diagnosis: biomarkers with downregulated expression, 2007b
5. Lawrence LaPointe, Susanne Pedersen, Glenn Brown, Lloyd Graham, and Graeme Young. A method of diagnosis: novel neoplasia marker (crng) with evidence of splice variants, 2007c

Bibliography

LA Aaltonen, P Peltomaki, FS Leach, P Sistonen, L Pylkkanen, JP Mecklin, H Jarvinen, SM Powell, J Jen, SR Hamilton, and al. et. Clues to the pathogenesis of familial colorectal cancer. *Science*, 260(5109):812–816, 1993.

H Aberle, A Bauer, J Stappert, A Kispert, and R Kemler. beta-catenin is a target for the ubiquitin-proteasome pathway. *EMBO J*, 16(13):3797–3804, 1997.

Jeroen Aerssens, Michael Camilleri, Willem Talloen, Leen Thielemans, Hinrich W H Gohlmann, Ilse Van Den Wyngaert, Theo Thielemans, Ronald De Hoogt, Christopher N Andrews, Adil E Bharucha, Paula J Carlson, Irene Busciglio, Duane D Burton, Thomas Smyrk, Raul Urrutia, and Bernard Coulie. Alterations in mucosal immunity identified in the colon of patients with irritable bowel syndrome. *Clin Gastroenterol Hepatol*, 6(2):194–205, 2008.

Affymetrix. *Statistical Algorithms Reference Guide*. Santa Clara, CA USA, 2001.

Affymetrix. *GeneChip Expression Analysis: Data Analysis Fundamentals*. Santa Clara, CA USA, 4th edition, 2004a.

Affymetrix. *Gene Expression Analysis: Technical Manual*. Santa Clara, CA USA, 2004b.

Affymetrix. *GeneChip Whole Transcript (WT) Sense Target Labeling Assay Manual*. Santa Clara, CA USA, 2007.

D Agrawal, T Chen, R Irby, J Quackenbush, AF Chambers, M Szabo, A Cantor, D Coppola, and TJ Yeatman. Osteopontin identified as lead marker of colon

cancer progression, using pooled sample expression profiling. *J Natl Cancer Inst*, 94(7):513–521, 2002.

A Aiello, E Tamborini, M Frattini, F Perrone, M Oggionni, S Pilotti, and MA Pierotti. Genetic markers in sporadic tumors. In MH Bronchud, MA foote, G giaccone, O olopade, and P workman, editors, *Principles of Molecular Oncology*, chapter 2. Humana Press Inc, Totowa, NJ, 2004.

William C Aird, Susan B Glueck, Victor J Dzau, and Richard E Pratt. Separating the wheat from the chaff: focus on "in silico data filtering to identify new angiogenesis targets from a large in vitro gene profile data set". *Physiol Genomics*, 10(1):1–3, 2002.

AA Alizadeh, DT Ross, CM Perou, and M van de Rijn. Towards a novel classification of human malignancies based on gene expression patterns. *J Pathol*, 195(1):41–52, 2001.

U Alon, N Barkai, DA Notterman, K Gish, S Ybarra, D Mack, and AJ Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci U S A*, 96(12):6745–6750, 1999.

Pedro M S Alves, Nicole Levy, Brian J Stevenson, Hanifa Bouzourene, Gregory Theiler, Gabriel Bricard, Sebastien Viatte, Maha Ayyoub, Henri Vuilleumier, Jean-Claude R Givel, Donata Rimoldi, Daniel E Speiser, C Victor Jongeneel, Pedro J Romero, and Frederic Levy. Identification of tumor-associated antigens by large-scale analysis of genes expressed in human colorectal cancer. *Cancer Immun*, 8(NIL):11, 2008.

Stefan Amatschek, Ulrich Koenig, Herbert Auer, Peter Steinlein, Margit Pacher, Agnes Gruenfelder, Gerhard Dekan, Sonja Vogl, Ernst Kubista, Karl-Heinz Heider, Christian Stratowa, Martin Schreiber, and Wolfgang Sommergruber. Tissue-wide expression profiling using cDNA subtraction and microarrays to identify tumor-specific genes. *Cancer Res*, 64(3):844–56, 2004.

Pauline Andreu, Sabine Colnot, Cecile Godard, Pierre Laurent-Puig, Dominique Lamarque, Axel Kahn, Christine Perret, and Beatrice Romagnolo. Identification of the IFITM family as a new molecular marker in human colorectal tumors. *Cancer Res*, 66(4):1949–55, 2006.

N Arber, H Hibshoosh, SF Moss, T Sutter, Y Zhang, M Begg, S Wang, IB Weinstein, and PR Holt. Increased expression of cyclin d1 is an early event in multistage colorectal carcinogenesis. *Gastroenterology*, 110(3):669–674, 1996.

LH Augenlicht, J Taylor, L Anderson, and M Lipkin. Patterns of gene expression that characterize the colonic mucosa in patients at genetic risk for colonic cancer. *Proc Natl Acad Sci U S A*, 88(8):3286–3289, 1991.

LH Augenlicht, JM Mariadason, A Wilson, D Arango, W Yang, BG Heerdt, and A Velcich. Short chain fatty acids and colon cancer. *J Nutr*, 132(12): 3804S–3808S, 2002.

Mark Ayers, Joseph Fargnoli, Anne Lewin, Qiuyan Wu, and J Suso Platero. Discovery and validation of biomarkers that respond to treatment with brivanib alaninate, a small-molecule VEGFR-2/FGFR-1 antagonist. *Cancer Res*, 67 (14):6899–906, 2007.

MW Babyatsky and DK Podolsky. Growth and development of the gastrointestinal tract. In Tadataka Yamada, DH Alpers, N Kaplowitz, L Laine, C Owyang, and DW Powell, editors, *Textbook of Gastroenterology, 4th Ed.*, pages 521–556. Lippincott Williams and Wilkins, Philadelphia, 2003.

S Backert, M Gelos, U Kobalz, ML Hanski, C Bohm, B Mann, N Lovin, A Gratchev, U Mansmann, MP Moyer, EO Riecken, and C Hanski. Differential gene expression in colon carcinoma cells and tissues detected with a cdna array. *Int J Cancer*, 82(6):868–874, 1999.

E Bair, T Hastie, P Debashis, and R Tibshirani. Prediction by supervised principal componentsvov. *J Am Statistical Assoc*, 101:119–137, 2006.

J Bara, J Nardelli, C Gadenne, M Prade, and P Burtin. Differences in the

expression of mucus-associated antigens between proximal and distal human colon adenocarcinomas. *Br J Cancer*, 49(4):495–501, 1984.

N Barker, A Hurlstone, H Musisi, A Miles, M Bienz, and H Clevers. The chromatin remodelling factor brg-1 interacts with beta-catenin to promote target gene activation. *EMBO J*, 20(17):4935–4943, 2001.

MD Bates, CR Erwin, LP Sanford, D Wiginton, JA Bezerra, LC Schatzman, AG Jegga, C Ley-Ebert, SS Williams, KA Steinbrecher, BW Warner, MB Cohen, and BJ Aronow. Novel genes and functional relationships in the adult mouse gastrointestinal tract identified by microarray analysis. *Gastroenterology*, 122(5):1467–1482, 2002.

R C Bates and A M Mercurio. The epithelial-mesenchymal transition (emt) and colorectal cancer progression. *Cancer Biol Ther*, 4(4):365–370, Apr 2005. URL `http://www.hubmed.org/display.cgi?uids=15846061`.

Richard C Bates and Arthur M Mercurio. Tumour necrosis factor-α stimulates the epithelial-to-mesenchymal transition fo human colonic organoids. *Mol Bio Cell*, 14, 2003.

E Batlle, JT Henderson, H Beghtel, MM van den Born, E Sancho, G Huls, J Meeldijk, J Robertson, M van de Wetering, T Pawson, and H Clevers. Beta-catenin and tcf mediate cell positioning in the intestinal epithelium by controlling the expression of ephb/ephrinb. *Cell*, 111(2):251–263, 2002.

S E Beck, B H Jung, A Fiorino, J Gomez, E D Rosario, B L Cabrera, S C Huang, J Y Chow, and J M Carethers. Bone morphogenetic protein signaling and growth suppression in colon cancer. *Am J Physiol Gastrointest Liver Physiol*, 291(1):135–145, Jul 2006. doi: 10.1152/ajpgi.00482.2005. URL `http://www.hubmed.org/display.cgi?uids=16769811`.

J Benhattar and E Saraga. Molecular genetics of dysplasia in ulcerative colitis. *Eur J Cancer*, 31A(7-8):1171–1173, 1995.

Y Benjamini and Y Hochberg. Controlling the false discovery rate: a practical

and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, 57, 1995.

E P Bennett, H Hassan, U Mandel, M A Hollingsworth, N Akisawa, Y Ikematsu, G Merkx, A G van Kessel, S Olofsson, and H Clausen. Cloning and characterization of a close homologue of human UDP-N-acetyl-alpha-D-galactosamine:Polypeptide N-acetylgalactosaminyltransferase-T3, designated GalNAc-T6. Evidence for genetic but not functional redundancy. *J Biol Chem*, 274(36):25362–70, 1999.

Nora Berois, Daniel Mazal, Luis Ubillos, Felipe Trajtenberg, Andre Nicolas, Xavier Sastre-Garau, Henri Magdelenat, and Eduardo Osinaga. UDP-N-acetyl-D-galactosamine: polypeptide N-acetylgalactosaminyltransferase-6 as a new immunohistochemical breast cancer marker. *J Histochem Cytochem*, 54 (3):317–28, 2006.

F Bertucci, S Salas, S Eysteries, V Nasser, P Finetti, C Ginestier, E Charafe-Jauffret, B Loriod, L Bachelart, J Montfort, G Victorero, F Viret, V Ollendorff, V Fert, M Giovaninni, JR Delpero, C Nguyen, P Viens, G Monges, D Birnbaum, and R Houlgatte. Gene expression profiling of colon cancer by dna microarrays and correlation with histoclinical parameters. *Oncogene*, 23 (7):1377–1391, 2004.

Yue-Hong Bian, Shu-Hong Huang, Ling Yang, Xiao-Li Ma, Jing-Wu Xie, and Hong-Wei Zhang. Sonic hedgehog-Gli1 pathway in colorectal adenocarcinomas. *World J Gastroenterol*, 13(11):1659–65, 2007.

Silvio Bicciato, Mario Pandin, Giuseppe Didone, and Carlo Di Bello. Pattern identification and classification in gene expression data using an autoassociative neural network model. *Biotechnol Bioeng*, 81(5):594–606, 2003.

M Bienz and H Clevers. Linking colorectal cancer to wnt signaling. *Cell*, 103(2) (2):311–320, 2000.

Andrea Bild and Phillip George Febbo. Application of a priori established gene sets to discover biologically important differential expression in microarray

data. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15278–15279, 2005. doi: 10.1073/pnas.0507477102. URL `http://www.pnas.org/content/102/43/15278.short`.

A P Bird. CpG-rich islands and the function of DNA methylation. *Nature*, 321 (6067):209–13, 1986.

K Birkenkamp-Demtroder, LL Christensen, SH Olesen, CM Frederiksen, P Laiho, LA Aaltonen, S Laurberg, FB Sorensen, R Hagemann, and TF ORntoft. Gene expression in colorectal cancer. *Cancer Res*, 62(15):4352–4363, 2002.

K Birkenkamp-Demtroder, SH Olesen, FB Sorensen, S Laurberg, P Laiho, LA Aaltonen, and TF Orntoft. Differential gene expression in colon cancer of the caecum versus the sigmoid and rectosigmoid. *Gut*, 54(3):374–384, 2005.

Chris Bishop. Training with noise is equivalent to tikhonov regularization. *Neural Comput*, 7(1):108–116, 1994.

Chris M Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, New York, 1st edition, 2006.

K Bkamp-Demtroder, LL Christensen, SH Olesen, CM Frederiksen, P Laiho, LA Aaltonen, S Laurberg, FB Sorensen, R Hagemann, and TF ORntoft. Gene expression in colorectal cancer. *Cancer Res*, 62(15):4352–4363, 2002.

J Martin Bland and Douglas G Altman. Statistics notes: Multiple significance tests: the Bonferroni method. *BMJ*, 310(6973):170–, 1995. URL `http://www.bmj.com`.

W F Bodmer, S Cottrell, A M Frischauf, I B Kerr, V A Murday, A J Rowan, M F Smith, E Solomon, H Thomas, and L Varesco. Genetic analysis of colorectal cancer. *Princess Takamatsu Symp*, 20(NIL):49–59, 1989.

CR Boland. Molecular genetics of hereditary nonpolyposis colorectal cancer. *Ann N Y Acad Sci*, 910:50–9; discussion 59–61, 2000.

C Bonithon-Kopp and AM Benhamiche. Are there several colorectal cancers? epidemiological data. *Eur J Cancer Prev*, 8 Suppl 1:S3–12, 1999.

C Booth and C S Potten. Gut instincts: thoughts on intestinal epithelial stem cells. *J Clin Invest*, 105(11):1493–1499, 2000. ISSN 0021-9738 (Print).

T Brabletz, A Jung, S Dag, F Hlubek, and T Kirchner. beta-catenin regulates the expression of the matrix metalloproteinase-7 in human colorectal cancer. *Am J Pathol*, 155(4):1033–1038, 1999.

L Breiman. Better subset selection using the non-negative garotte. *University of California Berkeley*, 1993.

MH Bronchud, MA Foote, G Giaccone, O Olopade, and P Workman. *Principles of Molecular Oncology, 2nd Ed.* Human Press, Totowa, NJ, 2004.

EB Brunschwig, K Wilson, D Mack, D Dawson, E Lawrence, JK Willson, S Lu, A Nosrati, RM Rerko, S Swinler, L Beard, JD Lutterbaugh, J Willis, P Platzer, and S Markowitz. Pmepa1, a transforming growth factor-beta-induced marker of terminal colonocyte differentiation whose expression is maintained in primary and metastatic colon cancer. *Cancer Res*, 63(7):1568–1575, 2003.

P Buckhaults, C Rago, B St Croix, KE Romans, S Saha, L Zhang, B Vogelstein, and KW Kinzler. Secreted and cell surface genes expressed in benign and malignant colorectal tumors. *Cancer Res*, 61(19):6996–7001, 2001.

P Buckhaults, Z Zhang, YC Chen, TL Wang, B St Croix, S Saha, A Bardelli, PJ Morin, K Polyak, RH Hruban, VE Velculescu, and IeM Shih. Identifying tumor origin using a gene expression-based classification map. *Cancer Res*, 63(14):4144–4149, 2003.

JA Bufill. Colorectal cancer: evidence for distinct genetic categories based on proximal or distal tumor location. *Ann Intern Med*, 113(10):779–788, 1990.

Chrisopher JC Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:121–167, 1998.

RW Burt. Familial risk and colon cancer. *Int J Cancer*, 69(1):44–46, 1996.

RW Burt and WS Samowitz. The adenomatous polyp and the hereditary polyposis syndromes. *Gastroenterol Clin North Am*, 17(4):657–678, 1988.

JC Byrd and RS Bresalier. Mucins and mucin binding proteins in colorectal cancer. *Cancer Metastasis Rev*, 23(1-2):77–99, 2004.

DP Cahill, C Lengauer, J Yu, GJ Riggins, JK Willson, SD Markowitz, KW Kinzler, and B Vogelstein. Mutations of mitotic checkpoint genes in human cancers. *Nature*, 392(6673):300–303, 1998.

G Calamita, A Mazzone, A Bizzoca, A Cavalier, G Cassano, D Thomas, and M Svelto. Expression and immunolocalization of the aquaporin-8 water channel in rat gastrointestinal tract. *Eur J Cell Biol*, 80(11):711–719, 2001.

J Caldero, E Campo, C Ascaso, J Ramos, MJ Panades, and JM Rene. Regional distribution of glycoconjugates in normal, transitional and neoplastic human colonic mucosa. a histochemical study using lectins. *Virchows Arch A Pathol Anat Histopathol*, 415(4):347–356, 1989.

Roger D Canales, Yuling Luo, James C Willey, Bradley Austermiller, Catalin C Barbacioru, Cecilie Boysen, Kathryn Hunkapiller, Roderick V Jensen, Charles R Knight, Kathleen Y Lee, Yunqing Ma, Botoul Maqsodi, Adam Papallo, Elizabeth Herness Peters, Karen Poulter, Patricia L Ruppel, Raymond R Samaha, Leming Shi, Wen Yang, Lu Zhang, and Federico M Goodsaid. Evaluation of DNA microarray results with quantitative gene expression platforms. *Nat Biotechnol*, 24(9):1115–22, 2006.

Massimiliano Cariati, Ali Naderi, John P Brown, Matthew J Smalley, Sarah E Pinder, Carlos Caldas, and Anand D Purushotham. Alpha-6 integrin is necessary for the tumourigenicity of a stem cell-like subpopulation within the MCF7 breast cancer cell line. *Int J Cancer*, 122(2):298–304, 2008.

AD Chalmers, JM Slack, and CW Beck. Regional gene expression in the epithelia of the xenopus tadpole gut. *Mech Dev*, 96(1):125–128, 2000.

AF Chambers and LM Matrisian. Changing views of the role of matrix metalloproteinases in metastasis. *J Natl Cancer Inst*, 89(17):1260–1270, 1997.

S K Chan, O L Griffith, I T Tai, and S J Jones. Meta-analysis of colorectal cancer gene expression profiling studies identifies consistently reported candidate biomarkers. *Cancer Epidemiol Biomarkers Prev*, 17(3):543–552, Mar 2008. doi: 10.1158/1055-9965.EPI-07-2615. URL `http://www.hubmed.org/display.cgi?uids=18349271`.

Guillaume Chatel, Corine Ganeff, Naima Boussif, Laurence Delacroix, Alexandra Briquet, Gregory Nolens, and Rosita Winkler. Hedgehog signaling pathway is inactive in colorectal cancer cell lines. *Int J Cancer*, 121(12):2622–7, 2007.

C Chatfield and A J Collins. *Introduction to Multivariate Analysis*. Chapman & Hall/CRC, 1981.

Yao Chen, Yi-Zeng Zhang, Zong-Guang Zhou, Gang Wang, and Zeng-Ni Yi. Identification of differently expressed genes in human colorectal adenocarcinoma. *World J Gastroenterol*, 12(7):1025–32, 2006.

Sou-Tyau Chiu, Fon-Jou Hsieh, Shi-Wen Chen, Chun-Lieh Chen, Hwei-Fan Shu, and Hung Li. Clinicopathologic correlation of up-regulated genes identified using cDNA microarray and real-time reverse transcription-PCR in human colorectal cancer. *Cancer Epidemiol Biomarkers Prev*, 14(2):437–43, 2005.

Dondapati Chowdary, Jessica Lathrop, Joanne Skelton, Kathleen Curtin, Thomas Briggs, Yi Zhang, Jack Yu, Yixin Wang, and Abhijit Mazumder. Prognostic gene expression signatures can be measured in tissues collected in RNAlater preservative. *J Mol Diagn*, 8(1):31–9, 2006.

PA Clarke, ML George, S Easdale, D Cunningham, RI Swift, ME Hill, DM Tait, and P Workman. Molecular pharmacology of cancer therapy in human colorectal cancer by gene expression profiling. *Cancer Res*, 63(20):6855–6863, 2003.

OR Colegio, CM Van Itallie, HJ McCrea, C Rahner, and JM Anderson. Claudins create charge-selective channels in the paracellular pathway between epithelial cells. *Am J Physiol Cell Physiol*, 283(1):C142–7, 2002.

Manuel Collado, Vanesa Garcia, Jose Miguel Garcia, Isabel Alonso, Luis Lombardia, Ramon Diaz-Uriarte, Luis A Lopez Fernandez, Angel Zaballos, Felix Bonilla, and Manuel Serrano. Genomic profiling of circulating plasma RNA for the analysis of cancer. *Clin Chem*, 53(10):1860–3, 2007.

Jason Comander, Sripriya Natarajan, Michael A Jr Gimbrone, and Guillermo Garcia-Cardena. Improving the statistical detection of regulated genes from microarray data using intensity-based variance estimation. *BMC Genomics*, 5(1):17, 2004.

ME Conacci-Sorrell, T Ben-Yedidia, M Shtutman, E Feinstein, P Einat, and A Ben-Ze'ev. Nr-cam is a target gene of the beta-catenin/lef-1 pathway in melanoma and colon cancer and its expression enhances motility and confers tumorigenesis. *Genes Dev*, 16(16):2058–2072, 2002.

Corinna Cortes and V Vapnik. Support vector networks. *Machine Learning*, 20: 273–297, 1995.

HC Crawford, BM Fingleton, LA Rudolph-Owen, KJ Goss, B Rubinfeld, P Polakis, and LM Matrisian. The metalloproteinase matrilysin is a target of beta-catenin transactivation in intestinal tumors. *Oncogene*, 18(18):2883–2891, 1999.

Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000.

RS Croner, K Guenther, T Foertsch, R Siebenhaar, WM Brueckl, C Stremmel, F Hlubek, W Hohenberger, and B Reingruber. Tissue preparation for gene expression profiling of colorectal carcinoma: three alternatives to laser microdissection with preamplification. *J Lab Clin Med*, 143(6):344–351, 2004.

JW Crott, SW Choi, JM Ordovas, JS Ditelberg, and JB Mason. Effects of dietary folate and aging on gene expression in the colonic mucosa of rats: implications for carcinogenesis. *Carcinogenesis*, 25(1):69–76, 2004.

MA Cuff, DW Lambert, and SP Shirazi-Beechey. Substrate-induced regulation of the human colonic monocarboxylate transporter, mct1. *J Physiol*, 539(Pt 2):361–371, 2002.

Schweinfest CW, KW Henderson, JR Gu, SD Kottaridis, S Besbeas, E Panotopoulou, and TS Papas. Subtraction hybridization cdna libraries from colon carcinoma and hepatic cancer. *Genet Anal Tech Appl*, 7(3):64–70, 1990.

Antonello D'Arrigo, Claudio Belluco, Alessandro Ambrosi, Maura Digito, Giovanni Esposito, Antonella Bertola, Michele Fabris, Valentina Nofrate, Enzo Mammano, Alberta Leon, Donato Nitti, and Mario Lise. Metastatic transcriptional pattern revealed by gene expression profiling in primary colorectal carcinoma. *Int J Cancer*, 115(2):256–62, 2005.

Susmita Datta and Somnath Datta. Empirical Bayes screening of many p-values with applications to microarray studies. *Bioinformatics*, 21(9):1987–94, 2005.

Mark M Davis, David I Cohen, Ellen A Nielsen, Michael Steinmetz, William E Paul, and Leroy Hood. Cell-type-specific cdna probes and the murine i region: The localization and orientation of a^d_α. *Proc Natl Acad Sci*, 81:2194–2198, 1984.

E Day. Is the periodic health examination worthwhile? *Cancer*, 47(5 Suppl): 1210–1214, 1981.

P de Santa Barbara, GR van den Brink, and DJ Roberts. Development and differentiation of the intestinal epithelium. *Cell Mol Life Sci*, 60(7):1322–1332, 2003.

G Deng, A Chen, J Hong, HS Chae, and YS Kim. Methylation of cpg in a small region of the hmlh1 promoter invariably correlates with the absence of gene expression. *Cancer Res*, 59(9):2029–2033, 1999.

G Deng, E Peng, J Gum, J Terdiman, M Sleisenger, and YS Kim. Methylation of hmlh1 promoter correlates with the gene silencing with a region-specific manner in colorectal cancer. *Br J Cancer*, 86(4):574–579, 2002.

BK Dieckgraefe, DL Crimmins, V Landt, C Houchen, S Anant, R Porche-Sorbet, and JH Ladenson. Expression of the regenerating gene family in inflammatory bowel disease mucosa: Reg ialpha upregulation, processing, and antiapoptotic activity. *J Investig Med*, 50(6):421–434, 2002.

P Distler and PR Holt. Are right- and left-sided colon neoplasms distinct tumors? *Dig Dis*, 15(4-5):302–311, 1997.

Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2000.

I Duluc, B Jost, and JN Freund. Multiple levels of control of the stage- and region-specific expression of rat intestinal lactase. *J Cell Biol*, 123(6 Pt 1): 1577–1586, 1993.

I Duluc, J N Freund, C Leberquier, and M Kedinger. Fetal endoderm primarily holds the temporal and positional information required for mammalian intestinal development. *J Cell Biol*, 126(1):211–221, 1994. ISSN 0021-9525 (Print).

Catherine I Dumur, Suhail Nasim, Al M Best, Kellie J Archer, Amy C Ladd, Valeria R Mas, David S Wilkinson, Carleton T Garrett, and Andrea Ferreira-Gonzalez. Evaluation of quality-control criteria for microarray gene expression analysis. *Clin Chem*, 50(11):1994–2002, 2004.

Bradley Efron and Robert Tibshirani. On testing the signficance of sets of genes. Technical report, Stanford Univeristy, Dept of Statistics, 2006.

B Ephrussi, RL Davidson, MC Weiss, H Harris, and G Klein. Malignancy of somatic cell hybrids. *Nature*, 224(226):1314–1316, 1969.

Anders Eriksson, Carl-Fredrik Flach, Anders Lindgren, Eva Kvifors, and Stefan Lange. Five mucosal transcripts of interest in ulcerative colitis identified by

quantitative real-time PCR: a prospective study. *BMC Gastroenterol*, 8(NIL): 34, 2008.

Steven Eschrich, Ivana Yang, Greg Bloom, Ka Yin Kwong, David Boulware, Alan Cantor, Domenico Coppola, Mogens Kruhoffer, Lauri Aaltonen, Torben F Orntoft, John Quackenbush, and Timothy J Yeatman. Molecular staging for survival prediction of colorectal cancer patients. *J Clin Oncol*, 23(15): 3526–35, 2005.

Julian J. Faraway. *Linear Models with R.* Chapman and Hall/CRC, 2004.

ER Fearon and B Vogelstein. A genetic model for colorectal tumorigenesis. *Cell*, 61(5):759–767, 1990.

MI Filipe and AC Branfoot. Mucin histochemistry of the colon. *Curr Top Pathol*, 63:143–178, 1976.

GG Finley, NT Schulz, SA Hill, JR Geiser, JM Pipas, and AI Meisler. Expression of the myc gene family in different stages of human colorectal cancer. *Oncogene*, 4(8):963–971, 1989.

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.

R Fodde, J Kuipers, C Rosenberg, R Smits, M Kielman, C Gaspar, JH van Es, C Breukel, J Wiegant, RH Giles, and H Clevers. Mutations in the apc tumour suppressor gene cause chromosomal instability. *Nat Cell Biol*, 3(4):433–438, 2001.

F Fogt, Z Zhuang, C Poremba, B Dockhorn-Dworniczak, and A Vortmeyer. Comparison of p53 immunoexpression with allelic loss of p53 in ulcerative colitis-associated dysplasia and carcinoma. *Oncol Rep*, 5(2):477–480, 1998.

C Foltzer-Jourdainne, M Kedinger, and F Raul. Perinatal expression of brush-border hydrolases in rat colon: hormonal and tissue regulations. *Am J Physiol*, 257(4 Pt 1):G496–503, 1989. ISSN 0002-9513 (Print).

CM Frederiksen, S Knudsen, S Laurberg, and TF Orntoft. Classification of dukes' b and c colorectal cancers using expression arrays. *J Cancer Res Clin Oncol*, 129(5):263–271, 2003.

Teresa Freire, Nora Berois, Cecilia Sonora, Mario Varangot, Enrique Barrios, and Eduardo Osinaga. UDP-N-acetyl-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 6 (ppGalNAc-T6) mRNA as a potential new marker for detection of bone marrow-disseminated breast cancer cells. *Int J Cancer*, 119(6):1383–8, 2006.

JB Friedman, EB Brunschwig, P Platzer, K Wilson, and SD Markowitz. C8orf4 is a transforming growth factor b induced transcript downregulated in metastatic colon cancer. *Int J Cancer*, 111(1):72–75, 2004.

M Fujita, Y Furukawa, Y Nagasawa, M Ogawa, and Y Nakamura. Down-regulation of monocyte chemotactic protein-3 by activated beta-catenin. *Cancer Res*, 60(23):6683–6687, 2000.

M Fujita, Y Furukawa, T Tsunoda, T Tanaka, M Ogawa, and Y Nakamura. Up-regulation of the ectodermal-neural cortex 1 (enc1) gene, a downstream target of the beta-catenin/t-cell factor complex, in colorectal carcinomas. *Cancer Res*, 61(21):7722–7726, 2001.

H Fukushima, H Yamamoto, F Itoh, S Horiuchi, Y Min, S Iku, and K Imai. Frequent alterations of the beta-catenin and tcf-4 genes, but not of the apc gene, in colon cancers with high-frequency microsatellite instability. *J Exp Clin Cancer Res*, 20(4):553–559, 2001.

GM Furnival and RM Wilson. Regressions by leaps and bounds. *Technometrics*, 16(4):499–511, 1974.

O Galamb, F Sipos, E Dinya, S Spisak, Z Tulassay, and B Molnar. mrna expression, functional profiling and multivariate classification of colon biopsy specimen by cdna overall glass microarray. *World J Gastroenterol*, 12(43): 6998–7006, 2006.

Orsolya Galamb, Ferenc Sipos, Norbert Solymosi, Sandor Spisak, Tibor Krenacs, Kinga Toth, Zsolt Tulassay, and Bela Molnar. Diagnostic mRNA expression patterns of inflamed, benign, and malignant colorectal biopsy specimen and their correlation with peripheral blood results. *Cancer Epidemiol Biomarkers Prev*, 17(10):2835–45, 2008.

J Garcia-Hirschfeld Garcia, A Blanes Berenguel, L Vicioso Recio, A Marquez Moreno, J Rubio Garrido, and A Matilla Vicente. Colon cancer: p53 expression and dna ploidy. their relation to proximal or distal tumor site. *Rev Esp Enferm Dig*, 91(7):481–488, 1999.

Laurent Gautier, Leslie Cope, Benjamin M Bolstad, and Rafael A Irizarry. affy–analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, 20 (3):307–15, 2004.

Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian Data Analysis*. Chapmon and Hall CRC, Boca Raton USA, 2nd edition, 2004.

GeneLogic. Review of data. Personal communication, 2005.

Robert C Gentleman, Vincent J. Carey, Douglas M. Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, Kurt Hornik, Torsten Hothorn, Wolfgang Huber, Stefano Iacus, Rafael Irizarry, Friedrich Leisch, Cheng Li, Martin Maechler, Anthony J. Rossini, Gunther Sawitzki, Colin Smith, Gordon Smyth, Luke Tierney, Jean Y. H. Yang, and Jianhua Zhang. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 5:R80, 2004. URL `http://genomebiology.com/2004/5/10/R80`.

E Georges-Labouesse, M Mark, N Messaddeq, and A Gansmuller. Essential role of alpha 6 integrins in cortical and retinal lamination. *Curr Biol*, 8(17):983–6, 1998.

ME Gerritsen, R Soriano, S Yang, G Ingle, C Zlot, K Toy, J Winer, A Draksharapu, F Peale, TD Wu, and PM Williams. In silico data filtering to identify

new angiogenesis targets from a large in vitro gene profiling data set. *Physiol Genomics*, 10(1):13–20, 2002.

F M Giardiello, S R Hamilton, A J Krush, S Piantadosi, L M Hylind, P Celano, S V Booker, C R Robinson, and G J Offerhaus. Treatment of colonic and rectal adenomas with sulindac in familial adenomatous polyposis. *N Engl J Med*, 328(18):1313–6, 1993.

R H Giles, J H van Es, and H Clevers. Caught up in a wnt storm: Wnt signaling in cancer. *Biochim Biophys Acta*, 1653(1):1–24, Jun 2003. URL `http://www.hubmed.org/display.cgi?uids=12781368`.

TJ Giordano, KA Shedden, DR Schwartz, R Kuick, JM Taylor, N Lee, DE Misek, JK Greenson, SL Kardia, DG Beer, G Rennert, KR Cho, SB Gruber, ER Fearon, and S Hanash. Organ-specific molecular classification of primary lung, colon, and ovarian adenocarcinomas using gene expression profiles. *Am J Pathol*, 159(4):1231–1238, 2001.

OK Glebov, LM Rodriguez, K Nakahara, J Jenkins, J Cliatt, CJ Humbyrd, J DeNobile, P Soballe, R Simon, G Wright, P Lynch, S Patterson, H Lynch, S Gallinger, A Buchbinder, G Gordon, E Hawk, and IR Kirsch. Distinguishing right from left colon by the pattern of gene expression. *Cancer Epidemiol Biomarkers Prev*, 12(8):755–762, 2003.

Ajay Goel, Takeshi Nagasaka, Christian N Arnold, Toru Inoue, Cody Hamilton, Donna Niedzwiecki, Carolyn Compton, Robert J Mayer, Richard Goldberg, Monica M Bertagnolli, and C Richard Boland. The CpG island methylator phenotype and chromosomal instability are inversely correlated in sporadic colorectal cancer. *Gastroenterology*, 132(1):127–38, 2007.

Gene Golub and F. Van Loan, Charles. *Matrix Computations, third ed.* The Johns Hopkins University Press, Baltimore, MD USA, 3rd edition, 1996. URL `www.press.jhu.edu`.

TR Golub, DK Slonim, P Tamayo, C Huard, M Gaasenbeek, JP Mesirov, H Coller, ML Loh, JR Downing, MA Caligiuri, CD Bloomfield, and ES Lan-

der. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, 1999.

JI Gordon and ML Hermiston. Differentiation and self-renewal in the mouse gastrointestinal epithelium. *Curr Opin Cell Biol*, 6(6):795–803, 1994.

Marian Grade, Patrick Hormann, Sandra Becker, Amanda B Hummon, Danny Wangsa, Sudhir Varma, Richard Simon, Torsten Liersch, Heinz Becker, Michael J Difilippantonio, B Michael Ghadimi, and Thomas Ried. Gene expression profiling reveals a massive, aneuploidy-dependent transcriptional deregulation and distinct differences between lymph node-negative and lymph node-positive colon carcinomas. *Cancer Res*, 67(1):41–56, 2007.

WM Grady. Genomic instability and colon cancer. *Cancer Metastasis Rev*, 23 (1-2):11–27, 2004.

F Guadagni, J Kantor, S Aloe, M D Carone, A Spila, R D'Alessandro, M R Abbolito, M Cosimelli, F Graziano, F Carboni, S Carlini, P Perri, F Sciarretta, J W Greiner, S V Kashmiri, S M Steinberg, M Roselli, and J Schlom. Detection of blood-borne cells in colorectal cancer patients by nested reverse transcription-polymerase chain reaction for carcinoembryonic antigen messenger RNA: longitudinal analyses and demonstration of its potential importance as an adjunct to multiple serum markers. *Cancer Res*, 61(6):2523–32, 2001.

JR Jr Gum, SC Crawley, JW Hicks, DE Szymkowski, and YS Kim. Muc17, a novel membrane-tethered mucin. *Biochem Biophys Res Commun*, 291(3): 466–475, 2002.

Jens K Habermann, Ulrike Paulsen, Uwe J Roblick, Madhvi B Upender, Lisa M McShane, Edward L Korn, Danny Wangsa, Stefan Kruger, Michael Duchrow, Hans-Peter Bruch, Gert Auer, and Thomas Ried. Stage-specific alterations of the genome, transcriptome, and proteome during colorectal carcinogenesis. *Genes Chromosomes Cancer*, 46(1):10–26, 2007.

SA Hahn, M Schutte, AT Hoque, CA Moskaluk, LT da Costa, E Rozenblum, CL Weinstein, A Fischer, CJ Yeo, RH Hruban, and SE Kern. Dpc4, a can-

didate tumor suppressor gene at human chromosome 18q21.1. *Science*, 271 (5247):350–353, 1996.

Mark Han, Choong Tsek Liew, Hong Wei Zhang, Samuel Chao, Run Zheng, Kok Thye Yip, Zhen-Ya Song, Hiu Ming Li, Xiao Ping Geng, Li Xin Zhu, Jian-Jiang Lin, K Wayne Marshall, and Choong Chin Liew. Novel blood-based, five-gene biomarker set for the detection of colorectal cancer. *Clin Cancer Res*, 14(2):455–60, 2008.

D. J Hand. *Construction and assessment of classification rules*, volume Construction and assessment of classification rules. Wiley, Chichester ; New York, 1997.

X Hao, M Du, AE Bishop, and IC Talbot. Imbalance between proliferation and apoptosis in the development of colorectal carcinoma. *Virchows Arch*, 433(6): 523–527, 1998.

JD Hardcastle, JO Chamberlain, MH Robinson, SM Moss, SS Amar, TW Balfour, PD James, and CM Mangham. Randomised controlled trial of faecal-occult-blood screening for colorectal cancer. *Lancet*, 348(9040):1472–1477, 1996.

Cesare Hassan, Emilio Di Giulio, Perry J Pickhardt, Angelo Zullo, Andrea Laghi, David H Kim, Franco Iafrate, and Sergio Morini. Cost effectiveness of colonoscopy, based on the appropriateness of an indication. *Clin Gastroenterol Hepatol*, 6(11):1231–6, 2008.

T Hastie and J Zhu. Classification of gene microarrays by penalized logistic regression. *Biostatistics*, 5(3):427–443, 2004.

T Hastie and J Zhu. Comment (to support vector machines with applications by moguerza and munoz). *Statistical Science*, 21(3):352–357, August 2006.

T Hastie, R Tibshirani, and J H Friedman. *The Elements of Statistical Learning*. Springer-Verlag, 2001.

T J Hastie and D Pregibon. *Statistical Models in S*, chapter Generalize linear models. Wadsworth & Brooks/Cole, 1992.

Trevor Hastie and Robert Tibshirani. Classification by pairwise coupling. In Michael I. Jordan, Michael J. Kearns, and Sara A. Solla, editors, *Advances in Neural Information Processing Systems*, volume 10. The MIT Press, 1998. URL `citeseer.nj.nec.com/hastie96classification.html`. Full version available at `http://www-stat.stanford.edu/~hastie/Papers/`.

Hiroko Hatano, Yasusei Kudo, Ikuko Ogawa, Takaaki Tsunematsu, Akira Kikuchi, Yoshimitsu Abiko, and Takashi Takata. IFN-induced transmembrane protein 1 promotes invasion at early stage of head and neck cancer progression. *Clin Cancer Res*, 14(19):6097–105, 2008.

N J Hawkins and R L Ward. Sporadic colorectal cancers with microsatellite instability and their possible origin in hyperplastic polyps and serrated adenomas. *J Natl Cancer Inst*, 93(17):1307–13, 2001.

Nicholas J Hawkins, Carolyn Bariol, and Robyn L Ward. The serrated neoplasia pathway. *Pathology*, 34(6):548–55, 2002.

TC He, AB Sparks, C Rago, H Hermeking, L Zawel, LT da Costa, PJ Morin, B Vogelstein, and KW Kinzler. Identification of c-myc as a target of the apc pathway. *Science*, 281(5382)(5382):1509–1512, 1998.

TC He, TA Chan, B Vogelstein, and KW Kinzler. Ppardelta is an apc-regulated target of nonsteroidal anti-inflammatory drugs. *Cell*, 99(3):335–345, 1999.

X C He, J Zhang, W G Tong, O Tawfik, J Ross, D H Scoville, Q Tian, X Zeng, X He, L M Wiedemann, Y Mishina, and L Li. Bmp signaling inhibits intestinal stem cell self-renewal through suppression of wnt-beta-catenin signaling. *Nat Genet*, 36(10):1117–1121, Oct 2004. doi: 10.1038/ng1430. URL `http://www.hubmed.org/display.cgi?uids=15378062`.

Steven C Hebert, David B Mount, and Gerardo Gamba. Molecular physiology of cation-coupled Cl- cotransport: the SLC12 family. *Pflugers Arch*, 447(5): 580–93, 2004.

JG Herman, A Umar, K Polyak, JR Graff, N Ahuja, JP Issa, S Markowitz, JK Willson, SR Hamilton, KW Kinzler, MF Kane, RD Kolodner, B Vogelstein, TA Kunkel, and SB Baylin. Incidence and functional consequences of hmlh1 promoter hypermethylation in colorectal carcinoma. *Proc Natl Acad Sci U S A*, 95(12):6870–6875, 1998.

MJ Hill, BC Morson, and HJ Bussey. Aetiology of adenoma–carcinoma sequence in large bowel. *Lancet*, 1(8058):245–247, 1978.

MO Hiltunen, L Alhonen, J Koistinaho, S Myohanen, M Paakkonen, S Marin, VM Kosma, and J Janne. Hypermethylation of the apc (adenomatous polyposis coli) gene promoter region in human colorectal carcinoma. *Int J Cancer*, 70(6):644–648, 1997.

AE Hoerl. Application of ridge analysis to regression problems. *Chemical Engineering Progress*, 58:54–59, 1962.

AE Hoerl and R Kennard. On regression analysis and biased estimation (abstract). *Technometrics*, 10:422–423, 1968.

AE Hoerl and R Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Yi Hong, Kok Sun Ho, Kong Weng Eu, and Peh Yean Cheah. A susceptibility gene set for early onset colorectal cancer that integrates diverse signaling pathways: implication for tumorigenesis. *Clin Cancer Res*, 13(4):1107–14, 2007.

SL Hostikka and MR Capecchi. The mouse hoxc11 gene: genomic structure and expression pattern. *Mech Dev*, 70(1-2):133–145, 1998.

R N Hourihan, G C O'Sullivan, and J G Morgan. Transcriptional gene expression profiles of oesophageal adenocarcinoma and normal oesophageal tissues. *Anticancer Res*, 23(1A):161–5, 2003.

K Hovanes, TW Li, JE Munguia, T Truong, T Milovanovic, J Lawrence Marsh, RF Holcombe, and ML Waterman. Beta-catenin-sensitive isoforms of lym-

phoid enhancer factor-1 are selectively expressed in colon cancer. *Nat Genet*, 28(1):53–57, 2001.

Peilin Huang, Jingmei Wang, Ying Guo, and Wei Xie. Molecular detection of disseminated tumor cells in the peripheral blood in patients with gastrointestinal cancer. *J Cancer Res Clin Oncol*, 129(3):192–8, 2003.

Yue Huang, Jun Fan, Jing Yang, and Guo-Zhang Zhu. Characterization of GPR56 protein and its suppressed expression in human pancreatic cancer cells. *Mol Cell Biochem*, 308(1-2):133–9, 2008.

Zhi Gang Huang, Zhi Hua Ran, Wei Lu, and Shu Don Xio. Analysis of gene expression profile in colon cancer using the cancer genome anatomy project and rna interfernce. *Chinese Journal of Digestive Diseases*, 7(NIL):97–102, 2006.

E W Hubbell, W M Liu, and R Mei. Robust estimators for expression analysis. *Bioinformatics*, 18:1585–1592, 2002.

M. Hubert, P. J. Rousseeuw, and K. Vanden Brandon. Robpca: a new approach to robust principal component analysis. *Technometrics*, 47:64–79, 2005.

B Iacopetta. Are there two sides to colorectal cancer? *Int J Cancer*, 101(5): 403–408, 2002.

John P Iaonnidis and Evangelia E Ntzani. Predictive ability of dna microarrays for cancer outcomes and correlates: an empirical assessment. *Lancet*, 362(1 November):1439–1444, 2003.

Yasushi Ichikawa, Takashi Ishikawa, Shinji Takahashi, Youhei Hamaguchi, Tomoyuki Morita, Itaru Nishizuka, Shigeki Yamaguchi, Itaru Endo, Hideyuki Ike, Shinji Togo, Shigeo Oki, Hiroshi Shimada, Koji Kadota, Shugo Nakamura, Hitoshi Goto, Hiroyuki Nitanda, Susumu Satomi, Takehito Sakai, Ichiei Narita, Fumitake Gejyo, Yasuhiro Tomaru, Kentaro Shimizu, Yoshihide Hayashizaki, and Yasushi Okazaki. Identification of genes regulating colorectal carcinogenesis by using the algorithm for diagnosing malignant state method. *Biochem Biophys Res Commun*, 296(2):497–506, 2002.

M Ilyas and IP Tomlinson. Genetic pathways in colorectal cancer. *Histopathology*, 28(5):389–399, 1996.

M Ilyas, IP Tomlinson, AM Hanby, T Yao, WF Bodmer, and IC Talbot. Bcl-2 expression in colorectal tumors: evidence of different pathways in sporadic and ulcerative-colitis-associated carcinomas. *Am J Pathol*, 149(5):1719–1726, 1996.

M Ilyas, JA Efstathiou, J Straub, HC Kim, and WF Bodmer. Transforming growth factor beta stimulation of colorectal cancer cell lines: type ii receptor bypass and changes in adhesion molecule expression. *Proc Natl Acad Sci U S A*, 96(6):3087–3091, 1999a.

M Ilyas, J Straub, IP Tomlinson, and WF Bodmer. Genetic pathways in colorectal and other cancers. *Eur J Cancer*, 35(14):1986–2002, 1999b.

Mohammad Ilyas, Jason Efstathiou, Josef Straub, He Kim, and Walter Bodmer. Transforming growth factor-β stimulation of colorectal cancer cell line: Type ii receptor bypass and adhesion molecule expression. *Proc Nat Acad Sci*, 96, 1999c.

John P Ioannidis. Microarrays and molecular research: noise discovery. *Lancet*, 365(5 Feb):454–455, 2005.

M Irigoyen, E Anso, E Salvo, J Dotor de las Herrerias, J J Martinez-Irujo, and A Rouzaut. TGFbeta-induced protein mediates lymphatic endothelial cell adhesion to the extracellular matrix under low oxygen conditions. *Cell Mol Life Sci*, 65(14):2244–55, 2008.

R W Irizarry, B M Bolstad, F Collin, L M Cope, B Hobbs, and T P Speed. Summaries of affymetrix genechip probe level data. *Nucleic Acid Research*, 31, 2003.

H Ishiguro, Y Furukawa, Y Daigo, Y Miyoshi, Y Nagasawa, T Nishiwaki, T Kawasoe, M Fujita, S Satoh, N Miwa, Y Fujii, and Y Nakamura. Isolation and characterization of human nbl4, a gene involved in the beta-catenin/tcf signaling pathway. *Jpn J Cancer Res*, 91(6):597–603, 2000.

R. James. Nucleic acid markers for use in determining predisposition to neoplasm and/or adenoma, 2001.

R James and J Kazenwadel. Homeobox gene expression in the intestinal epithelium of adult mice. *J. Biol. Chem.*, 266(5):3246–3251, 1991. URL `http://www.jbc.org/cgi/content/abstract/266/5/3246`.

R James and J Kazenwadel. Unpublished differential display research. Unpublished manuscript, 2002.

R James, T Erler, and J Kazenwadel. Structure of the murine homeobox gene cdx-2. expression in embryonic and adult intestinal epithelium. *J Biol Chem*, 269(21):15229–15237, 1994.

E Jansova, I Koutna, P Krontorad, Z Svoboda, S Krivankova, J Zaloudik, M Kozubek, and S Kozubek. Comparative transcriptome maps: a new approach to the diagnosis of colorectal carcinoma patients using cDNA microarrays. *Clin Genet*, 69(3):218–27, 2006.

J R Jass. Classification of colorectal cancer based on correlation of clinical, morphological and molecular features. *Histopathology*, 50(1):113–30, 2007a.

Jeremy R Jass. Serrated adenoma of the colorectum and the DNA-methylator phenotype. *Nat Clin Pract Oncol*, 2(8):398–405, 2005.

Jeremy R Jass. Molecular heterogeneity of colorectal cancer: Implications for cancer control. *Surg Oncol*, 16 Suppl 1(NIL):S7–9, 2007b.

B Jeansonne, Q Lu, DA Goodenough, and YH Chen. Claudin-8 interacts with multi-pdz domain protein 1 (mupp1) and reduces paracellular conductance in epithelial cells. *Cell Mol Biol (Noisy-le-grand)*, 49(1):13–21, 2003.

AM Jubb, TQ Pham, AM Hanby, GD Frantz, FV Peale, TD Wu, HW Koeppen, and KJ Hillan. Expression of vascular endothelial growth factor, hypoxia inducible factor 1alpha, and carbonic anhydrase ix in human tumours. *J Clin Pathol*, 57(5):504–512, 2004.

R Kalaba, R Xu, and W Feng. Solving shortest length least squares problems via dynamic programming. *J Optimization Theory and Application*, 85(3): 613–632, 1995.

Minoru Kanehisa and Susumu Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucl. Acids Res.*, 28(1):27–30, 2000. doi: 10.1093/nar/28.1.27.

Minoru Kanehisa, Michihiro Araki, Susumu Goto, Masahiro Hattori, Mika Hirakawa, Masumi Itoh, Toshiaki Katayama, Shuichi Kawashima, Shujiro Okuda, Toshiaki Tokimatsu, and Yoshihiro Yamanishi. Kegg for linking genomes to life and the environment. *Nucl Acids Res*, 36(suppl1):D480–484, 2008. doi: 10.1093/nar/gkm882.

KB Kaplan, AA Burds, JR Swedlow, SS Bekir, PK Sorger, and IS Nathke. A role for the adenomatous polyposis coli protein in chromosome segregation. *Nat Cell Biol*, 3(4):429–432, 2001.

Masuko Katoh and Masaru Katoh. Notch signaling in gastrointestinal tract (review). *Int J Oncol*, 30(1):247–51, 2007.

Yuriko Katoh and Masaru Katoh. Hedgehog signaling pathway and gastrointestinal stem cell signaling network (review). *Int J Mol Med*, 18(6):1019–23, 2006.

H Kawasaki, M Toyoda, H Shinohara, J Okuda, I Watanabe, T Yamamoto, K Tanaka, T Tenjo, and N Tanigawa. Expression of survivin correlates with apoptosis, proliferation, and angiogenesis during human colorectal tumorigenesis. *Cancer*, 91(11):2026–2032, 2001.

T Kawasoe, Y Furukawa, Y Daigo, T Nishiwaki, H Ishiguro, M Fujita, S Satoh, N Miwa, Y Nagasawa, Y Miyoshi, M Ogawa, and Y Nakamura. Isolation and characterization of a novel human gene, drctnnb1a, the expression of which is down-regulated by beta-catenin. *Cancer Res*, 60(13):3354–3358, 2000.

M Kedinger, O Lefebvre, I Duluc, J N Freund, and P Simon-Assmann. Cellular and molecular partners involved in gut morphogenesis and differentiation.

Philos Trans R Soc Lond B Biol Sci, 353(1370):847–856, 1998. ISSN 0962-8436 (Print). doi: 10.1098/rstb.1998.0249.

W Kemmner, C Roefzaad, W Haensch, and PM Schlag. Glycosyltransferase expression in human colonic tissue examined by oligonucleotide arrays. *Biochim Biophys Acta*, 1621(3):272–279, 2003.

H Kiiveri, Robert Dunne, and Lawrence LaPointe. Canonical variate analysis and microarrays. In *AMATA 2005 Meeting*, 2005. Poster: AMATA (Adelaide).

H T Kiiveri. Canonical variate analysis of high–dimensional spectral data. *Technometrics*, 34(3):321–331, 1992.

IJ Kim, HC Kang, JH Park, Y Shin, JL Ku, SB Lim, SY Park, SY Jung, HK Kim, and JG Park. Development and applications of a beta-catenin oligonucleotide microarray: beta-catenin mutations are dominantly found in the proximal colon cancers with microsatellite instability. *Clin Cancer Res*, 9 (8):2920–2925, 2003a.

Jin-Cheon Kim, Seon-Young Kim, Seon-Ae Roh, Dong-Hyung Cho, Dae-Dong Kim, Jeong-Hyun Kim, and Yong-Sung Kim. Gene expression profiling: Canonical molecular changes and clinicopathological features in sporadic colorectal cancers. *World J Gastroenterol*, 14(43):6662–72, 2008a.

JS Kim, H Crooks, T Dracheva, TG Nishanian, B Singh, J Jen, and T Waldman. Oncogenic beta-catenin is required for bone morphogenetic protein 4 expression in human cancer cells. *Cancer Res*, 62(10):2744–2748, 2002.

Kyongrae Kim, Ungchae Park, Joonho Wang, Jaedong Lee, Seunghwa Park, Sangyoon Kim, Dongkug Choi, Changil Kim, and Jiyoung Park. Gene profiling of colonic serrated adenomas by using oligonucleotide microarray. *Int J Colorectal Dis*, 23(6):569–80, 2008b.

PJ Kim, J Plescia, H Clevers, ER Fearon, and DC Altieri. Survivin and molecular pathogenesis of colorectal cancer. *Lancet*, 362(9379):205–209, 2003b.

KW Kinzler and B Vogelstein. Lessons from hereditary colorectal cancer. *Cell*, 87(2):159–170, 1996.

KW Kinzler and B Vogelstein. Cancer-susceptibility genes. gatekeepers and caretakers. *Nature*, 386(6627):761, 763, 1997.

O Kitahara, Y Furukawa, T Tanaka, C Kihara, K Ono, R Yanagawa, ME Nita, T Takagi, Y Nakamura, and T Tsunoda. Alterations of gene expression during colorectal carcinogenesis revealed by cdna microarrays after laser-capture microdissection of tumor tissues and normal epithelia. *Cancer Res*, 61(9): 3544–3549, 2001.

A Klaus and W Birchmeier. Wnt signalling and its impact on development and cancer. *Nat Rev Cancer*, 8(5):387–398, May 2008. doi: 10.1038/nrc2389. URL `http://www.hubmed.org/display.cgi?uids=18432252`.

AG Knudson. Mutation and human cancer. *Adv Cancer Res*, 17:317–352, 1973.

AG Knudson. Antioncogenes and human cancer. *Proc Natl Acad Sci U S A*, 90 (23):10914–10921, 1993.

AG Jr Knudson. Mutation and cancer: statistical study of retinoblastoma. *Proc Natl Acad Sci U S A*, 68(4):820–823, 1971.

U Koch and F Radtke. Notch and cancer: a double-edged sword. *Cell Mol Life Sci*, 64(21):2746–62, 2007.

Astrid Koehler, Frauke Bataille, Cornelia Schmid, Petra Ruemmele, Annette Waldeck, Hagen Blaszyk, Arndt Hartmann, Ferdinand Hofstaedter, and Wolfgang Dietmaier. Gene expression profiling of colorectal cancer and metastases divides tumours according to their clinicopathological stage. *J Pathol*, 204(1): 65–74, 2004.

TJ Koh, CJ Bulitta, JV Fleming, GJ Dockray, A Varro, and TC Wang. Gastrin is a target of the beta-catenin/tcf-4 growth-signaling pathway in a model of intestinal polyposis. *J Clin Invest*, 106(4):533–539, 2000.

FT Kolligs, MT Nieman, I Winer, G Hu, D Van Mater, Y Feng, IM Smith, R Wu, Y Zhai, KR Cho, and ER Fearon. Itf-2, a downstream target of the wnt/tcf pathway, is activated in human cancers with beta-catenin defects and promotes neoplastic transformation. *Cancer Cell*, 1(2):145–155, 2002.

K Komuro, M Tada, E Tamoto, A Kawakami, A Matsunaga, K Teramoto, G Shindoh, M Takada, K Murakawa, M Kanai, N Kobayashi, Y Fujiwara, N Nishimura, J Hamada, A Ishizu, H Ikeda, S Kondo, H Katoh, T Moriuchi, and T Yoshiki. Right- and left-sided colorectal cancers display distinct expression profiles and the anatomical stratification allows a high accuracy prediction of lymph node metastasis. *J Surg Res*, 124(2):216–224, 2005.

T Kondo, P Dolle, J Zakany, and D Duboule. Function of posterior hoxd genes in the morphogenesis of the anal sphincter. *Development*, 122(9):2651–2659, 1996.

V Korinek, N Barker, PJ Morin, D van Wichen, R de Weger, KW Kinzler, B Vogelstein, and H Clevers. Constitutive transcriptional activation by a beta-catenin-tcf complex in apc-/- colon carcinoma. *Science*, 275(5307):1784–1787, 1997.

K Kosaki, R Kosaki, T Suzuki, H Yoshihashi, T Takahashi, K Sasaki, M Tomita, W McGinnis, and N Matsuo. Complete mutation analysis panel of the 39 human hox genes. *Teratology*, 65(2):50–62, 2002.

O Kronborg, C Fenger, J Olsen, OD Jorgensen, and O Sondergaard. Randomised study of screening for colorectal cancer with faecal-occult-blood test. *Lancet*, 348(9040):1467–1471, 1996.

W.J. Krzanowski and F.H.C. Marriott. *Multivariate Analysis: Classification, Covariance Structures and Repeated Measurements Pt. 2 (Kendall's Library of Statistics)*. Hodder Arnold, 1995.

H.W. Kuhn and A.W. Tucker. Nonlinear programming. In J. Neyman, editor, *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pages 481–492. University of California Press, 1951.

G Kusakai, A Suzuki, T Ogura, S Miyamoto, A Ochiai, M Kaminishi, and H Esumi. Ark5 expression in colorectal cancer and its implications for tumor progression. *Am J Pathol*, 164(3):987–995, 2004.

Hyuk-Chan Kwon, Sung-Hyun Kim, Mee-Sook Roh, Jae-Seok Kim, Hyung-Sik Lee, Hong-Jo Choi, Jin-Sook Jeong, Hyo-Jin Kim, and Tae-Ho Hwang. Gene expression profiling in lymph node-positive and lymph node-negative colorectal cancer. *Dis Colon Rectum*, 47(2):141–52, 2004.

Sofia Lagerholm, Sara Lagerholm, Sudhir Dutta, and Padmanabhan Nair. Non-invasive detection of c-myc p64, c-myc p67 and c-erbb-2 in colorectal cancer. *Scand J Gastroenterol*, 40(11):1343–50, 2005.

Lawrence LaPointe. The normal colon gene map: from maths to genes. In *Australian Gastroenterology Week*, Adelaide, 2006. Invited Session Presentation.

Lawrence LaPointe. Biomarkers for colorectal neoplasia. In *M D Anderson Cancer Center*, Houston, TX USA, 2007a. Invited Seminar.

Lawrence LaPointe. Brave new world: Advances in genomics; gene expression mapping of the normal colon. In *New Zealand Bio*, Auckland, NZ, 2007b. Invited Session Presentation.

Lawrence LaPointe and Robert Dunne. Quality control analysis of genelogic data. Technical Report 05/2005, CSIRO, Preventative Health Flagship, 2005a.

Lawrence LaPointe and Robert Dunne. Identification of colorectal cancer biomarkers using publicy available gene expression data. In *Gastroenterology*. AGA, 2005b. Poster: DDW (Chicago).

Lawrence LaPointe and Robert Dunne. Comparison of machine learning techniques to identify biomarkers for colorectal cancer in publicly available data. In *International Society of Computational Biology*, 2005c. Poster: ISMB (Detroit, USA).

Lawrence LaPointe and Robert Dunne. A method of diagnosis: markers of anatomical location, 2005d.

Lawrence LaPointe and Robert Dunne. Normalization of custom microarrays. In *AMATA 2007 Meeting*, 2007. Poster: AMATA (Brisbane).

Lawrence LaPointe, Daniel Worthely, and Robert Dunne. Comparison of classification methods to discover biomarkers based on gene expression. In *Conference of Australian Microarrays and Associated Technologies (AMATA)*, 2005a.

Lawrence LaPointe, Graeme P Young, and Howard Chandler. Analysis of mrna expression profiles in colorectal adenomas using k-nearest neighbor cluster analysis. In *Gastroenterology*. AGA, 2005b. Poster: DDW (Orlando).

Lawrence LaPointe, R Dunne, G Young, T Lockett, B Wilson, and P Molloy. Nucleic acid markers for use in determining predisposition to neoplasm and/or adenoma, 2007a.

Lawrence LaPointe, Robert Dunne, Graeme Young, Peter Molloy, Trevor Lockett, and William Wilson. A method of diagnosis: biomarkers with downregulated expression, 2007b.

Lawrence LaPointe, Susanne Pedersen, Glenn Brown, Lloyd Graham, and Graeme Young. A method of diagnosis: novel neoplasia marker (crng) with evidence of splice variants, 2007c.

Lawrence LaPointe, Robert Dunne, Glenn S Brown, Daniel L Worthley, Peter L. Molloy, David Wattchow, and Graeme P. Young. Map of differential transcript expression in the normal human large intestine. *Physiol. Genomics*, 33(1):50–64, 2008.

Lawrence LaPointe, Robert Dunne, Peter Molloy, L Clark, Thu Ho, Susanne Pedersen, and Graeme P Young. Biomarkers with high sensitivity and specicity for colorectal adenomas and carcinomas. In *Gastroenterology*. AGA, 2009. Poster: DDW (Chicago).

PJ Laybourn and JT Kadonaga. Threshold phenomena and long-distance activation of transcription by rna polymerase ii. *Science*, 257(5077):1682–1685, 1992.

S Lechner, U Muller-Ladner, E Neumann, W Dietmaier, J Welsh, J Scholmerich, J Ruschoff, and F Kullmann. Use of simplified transcriptors for the analysis of gene expression profiles in laser-microdissected cell populations. *Lab Invest*, 81(9):1233–1242, 2001.

Sam W Lee, Catherine Tamasetto, and Ruth Sager. Positive selection of candidate tumor-suppressor genes by subtractive hybridization. *Proc Natl Acad Sci*, 88:2825–2829, 1991.

C Lengauer, KW Kinzler, and B Vogelstein. Genetic instability in colorectal cancers. *Nature*, 386(6625):623–627, 1997.

Bernard Levin, David A Lieberman, Beth McFarland, Kimberly S Andrews, Durado Brooks, John Bond, Chiranjeev Dash, Francis M Giardiello, Seth Glick, David Johnson, C Daniel Johnson, Theodore R Levin, Perry J Pickhardt, Douglas K Rex, Robert A Smith, Alan Thorson, and Sidney J Winawer. Screening and surveillance for the early detection of colorectal cancer and adenomatous polyps, 2008: a joint guideline from the American Cancer Society, the US Multi-Society Task Force on Colorectal Cancer, and the American College of Radiology. *Gastroenterology*, 134(5):1570–95, 2008.

L Li, TA Darden, CR Weinberg, AJ Levine, and LG Pedersen. Gene assessment and sample classification for gene expression data using a genetic algorithm/k-nearest neighbor method. *Comb Chem High Throughput Screen*, 4(8):727–739, 2001a.

Leping Li, Clarice R Weinberg, Thomas A Darden, Lee G Pedersen, and LG Pedersen. Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the ga/knn method. *Bioinformatics*, 17(12):1131–1142, 2001b.

Zejuan Li, Roger T Luo, Suangli Mi, Miao Sun, Ping Chen, Jingye Bao, Mary Beth Neilly, Nimanthi Jayathilaka, Deborah S Johnson, Lili Wang, Jun Yu, Huanming Yang, San Ming Wang, Janet Rowley, Jianjun Chen, and Michael Thirman. Consistent Deregulation of Gene Expression between Hu-

man and Murine MLL Rearrangement Leukemias. *Cancer Res*, NIL(NIL): NIL, 2009.

John J Liang, Sadir Alrawi, and Dongfeng Tan. Nomenclature, molecular genetics and clinical significance of the precursor lesions in the serrated polyp pathway of colorectal carcinoma. *Int J Clin Exp Pathol*, 1(4):317–24, 2008.

Peng Liang and Arthur B Pardee. Differential display of eukaryotic messenger rna by means fo the polymerase chain reaction. *Science*, 257(5072):967–971, 1992.

David Lieberman. Colonoscopy: as good as gold? *Ann Intern Med*, 141(5): 401–3, 2004.

GJ Liefers and RA Tollenaar. Cancer genetics and their application to individualised medicine. *Eur J Cancer*, 38(7):872–879, 2002.

LJ Lin, CQ Zheng, Y Jin, Y Ma, WG Jiang, and T Ma. Expression of survivin protein in human colorectal carcinogenesis. *World J Gastroenterol*, 9(5):974–977, 2003.

YM Lin, K Ono, S Satoh, H Ishiguro, M Fujita, N Miwa, T Tanaka, T Tsunoda, KC Yang, Y Nakamura, and Y Furukawa. Identification of af17 as a downstream gene of the beta-catenin/t-cell factor pathway and its involvement in colorectal carcinogenesis. *Cancer Res*, 61(17):6345–6349, 2001.

YM Lin, Y Furukawa, T Tsunoda, CT Yue, KC Yang, and Y Nakamura. Molecular diagnosis of colorectal tumors by expression profiles of 50 genes expressed differentially in adenomas and carcinomas. *Oncogene*, 21(26):4120–4128, 2002.

Lance A Liotta and Elise C Kohn. The microenvironment of the tumour-host interface. *Nature*, 411(17 May):375–379, 2001.

RJ Lipshutz, SP Fodor, TR Gingeras, and DJ Lockhart. High density synthetic oligonucleotide arrays. *Nat Genet*, 21(1 Suppl):20–24, 1999.

B Liu, R Parsons, N Papadopoulos, NC Nicolaides, HT Lynch, P Watson, JR Jass, M Dunlop, A Wyllie, P Peltomaki, A de la Chapelle, SR Hamilton, B Vogelstein, and KW Kinzler. Analysis of mismatch repair genes in hereditary non-polyposis colorectal cancer patients. *Nat Med*, 2(2):169–174, 1996.

M Liu, R M Parker, K Darby, H J Eyre, N G Copeland, J Crawford, D J Gilbert, G R Sutherland, N A Jenkins, and H Herzog. GPR56, a novel secretin-like human G-protein-coupled receptor gene. *Genomics*, 55(3):296–305, 1999.

Ting-Yuan Liu, ChenWei Lin, Seth Falcon, Jianhua JZhang, and James W MacDonald. Kegg data package. WWW, 2008a. URL `http://www.bioconductor.org/packages/2.2/data/annotation/html/KEGG.html`.

XF Liu, P Olsson, CD Wolfgang, TK Bera, P Duray, B Lee, and I Pastan. Prac: A novel small nuclear protein that is specifically expressed in human prostate and colon. *Prostate*, 47(2):125–131, 2001.

Yu-Hu Liu, Jua Lin, Jian Guo, Zhi-Jian You, Zai-Guo Wang, Dong Zhong, Xing-Long Yang, Zhen-Shu Zhang, Bing Xiao, and Wen-Ying Guo. [Detection of interferon-induced transmembrane-1 gene expression for clinical diagnosis of colorectal cancer.]. *Nan Fang Yi Ke Da Xue Xue Bao*, 28(11):1950–3, 2008b.

Alexandre Loktionov. Cell exfoliation in the human colon: myth, reality and implications for colorectal cancer screening. *Int J Cancer*, 120(11):2281–2289, 2007. ISSN 0020-7136 (Print). doi: 10.1002/ijc.22647.

T A Longacre and C M Fenoglio-Preiser. Mixed hyperplastic adenomatous polyps/serrated adenomas. A distinct form of colorectal neoplasia. *Am J Surg Pathol*, 14(6):524–37, 1990.

E Lucci-Cordisco, I Zito, F Gensini, and M Genuardi. Hereditary nonpolyposis colorectal cancer and related conditions. *Am J Med Genet*, 122A(4):325–334, 2003.

W m Liu, R Mei, X Di, T B Ryder, E Hubbell, S Dee, T A Webster, C A Harrington, M h Ho, J Baid, and S P Smeekens. Analysis of high density

expression microarrays with signed-rank call algorithms. *Bioinformatics*, 18 (12):1593–9, 2002.

Chaoyu Ma, Yu Rong, Daniel R Radiloff, Michael B Datto, Barbara Centeno, Shideng Bao, Anthony Wai Ming Cheng, Fumin Lin, Shibo Jiang, Timothy J Yeatman, and Xiao-Fan Wang. Extracellular matrix protein betaig-h3/TGFBI promotes metastasis of colon cancer by enhancing cell extravasation. *Genes Dev*, 22(3):308–21, 2008.

GT Macfarlane, GR Gibson, and JH Cummings. Comparison of fermentation reactions in different regions of the human colon. *J Appl Bacteriol*, 72(1): 57–64, 1992.

Blair B Madison, Katherine Braunstein, Erlene Kuizon, Kathleen Portman, Xiaotan T Qiao, and Deborah L Gumucio. Epithelial hedgehog signals pattern the intestinal crypt-villus axis. *Development*, 132(2):279–89, 2005.

R Maglietta, A Piepoli, D Catalano, F Licciulli, M Carella, S Liuni, G Pesole, F Perri, and N Ancona. Statistical assessment of functional categories of genes deregulated in pathological conditions by using microarray data. *Bioinformatics*, 23(16):2063–72, 2007.

A Malliri, WA Yeudall, M Nikolic, DH Crouch, EK Parkinson, and B Ozanne. Sensitivity to transforming growth factor beta 1-induced growth arrest is common in human squamous cell carcinoma cell lines: c-myc down-regulation and p21waf1 induction are important early events. *Cell Growth Differ*, 7(10): 1291–1304, 1996.

JS Mandel, JH Bond, TR Church, DC Snover, GM Bradley, LM Schuman, and F Ederer. Reducing mortality from colorectal cancer by screening for fecal occult blood. minnesota colon cancer control study. *N Engl J Med*, 328(19): 1365–1371, 1993.

JS Mandel, TR Church, JH Bond, F Ederer, MS Geisser, SJ Mongin, DC Snover, and LM Schuman. The effect of fecal occult-blood screening on the incidence of colorectal cancer. *N Engl J Med*, 343(22):1603–1607, 2000.

B Mann, M Gelos, A Siedow, ML Hanski, A Gratchev, M Ilyas, WF Bodmer, MP Moyer, EO Riecken, HJ Buhr, and C Hanski. Target genes of beta-catenin-t cell-factor/lymphoid-enhancer-factor signaling in human colorectal carcinomas. *Proc Natl Acad Sci U S A*, 96(4):1603–1608, 1999.

JM Mariadason, D Arango, GA Corner, MJ Aranes, KA Hotchkiss, W Yang, and LH Augenlicht. A gene expression profile that defines colon cell maturation in vitro. *Cancer Res*, 62(16):4791–4804, 2002.

AJ Markowitz and SJ Winawer. Screening and surveillance for colorectal cancer. *Semin Oncol*, 26(5):485–498, 1999.

Emma Marshman, Catherine Booth, and Christopher S Potten. The intestinal epithelial stem cell. *Bioessays*, 24(1):91–98, 2002. ISSN 0265-9247 (Print). doi: 10.1002/bies.10028.

J Massague, SW Blain, and RS Lo. Tgfbeta signaling in growth control, cancer, and heritable disorders. *Cell*, 103(2):295–309, 2000.

H Masuda, Y Takahashi, S Asai, and T Takayama. Distinct gene expression of osteopontin in patients with ulcerative colitis. *J Surg Res*, 111(1):85–90, 2003.

Kaoichi Matsuzaki and Kazuichi Okazaki. Transforming growth factor-β during carcinogenesis: the shift from epithelial to mesenchymal signaling. *Gastroenterology*, 41, 2006.

CM McIver, JM Lloyd, PJ Hewett, and JE Hardingham. Dipeptidase 1: a candidate tumor-specific molecular marker in colorectal carcinoma. *Cancer Lett*, 209(1):67–74, 2004.

Geoffrey McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley and Sons, New Jersey USA, 1st edition, 1992.

Christian Melle, Gunther Ernst, Bettina Schimmel, Annett Bleul, and Ferdinand von Eggeling. Colon-derived liver metastasis, colorectal carcinoma, and hepatocellular carcinoma can be discriminated by the Ca(2+)-binding proteins S100A6 and S100A11. *PLoS ONE*, 3(12):e3767, 2008.

D Mennerich, A Vogel, I Klaman, E Dahl, RB Lichtner, A Rosenthal, HD Pohlenz, KH Thierauch, and A Sommer. Shift of syndecan-1 expression from epithelial to stromal cells during progression of solid tumours. *Eur J Cancer*, 40(9):1373–1382, 2004.

M Michael. Review of tissue sample annotation and possible miscoding. Personal communication, 2008.

GL Miklos and R Maleszka. Microarray reality checks in the context of a complex disease. *Nat Biotechnol*, 22(5):615–621, 2004.

JR Miller. The wnts. *Genome Biol*, 3(1)(1):REVIEWS3001, 2002.

N Miwa, M Furuse, S Tsukita, N Niikawa, Y Nakamura, and Y Furukawa. Involvement of claudin-1 in the beta-catenin/tcf signaling pathway and its frequent upregulation in human colorectal cancers. *Oncol Res*, 12(11-12):469–476, 2000.

MM Mogensen, JB Tucker, JB Mackie, AR Prescott, and IS Nathke. The adenomatous polyposis coli protein unambiguously localizes to microtubule plus ends and is involved in establishing parallel arrays of microtubule bundles in highly polarized epithelial cells. *J Cell Biol*, 157(6):1041–1048, 2002.

JM Moguerza and A Munoz. Support vector machines with applications. *Statistical Science*, 21(3):322–336, August 2006.

RK Montgomery, AE Mulberg, and RJ Grand. Development of the human gastrointestinal tract: twenty years of progress. *Gastroenterology*, 116(3):702–731, 1999.

Daisuke Mori, Yuji Nakafusa, Kohji Miyazaki, and Osamu Tokunaga. Differential expression of Janus kinase 3 (JAK3), matrix metalloproteinase 13 (MMP13), heat shock protein 60 (HSP60), and mouse double minute 2 (MDM2) in human colorectal cancer progression using human cancer cDNA microarrays. *Pathol Res Pract*, 201(12):777–89, 2005.

Y Mori, FM Selaru, F Sato, J Yin, LA Simms, Y Xu, A Olaru, E Deacu, S Wang, JM Taylor, J Young, B Leggett, JR Jass, JM Abraham, D Shibata, and SJ Meltzer. The impact of microsatellite instability on the molecular phenotype of colorectal tumors. *Cancer Res*, 63(15):4577–4582, 2003.

Y Mori, J Yin, F Sato, A Sterian, LA Simms, FM Selaru, K Schulmann, Y Xu, A Olaru, S Wang, E Deacu, JM Abraham, J Young, BA Leggett, and SJ Meltzer. Identification of genes uniquely involved in frequent microsatellite instability colon carcinogenesis by expression profiling combined with epigenetic scanning. *Cancer Res*, 64(7):2434–2438, 2004.

PJ Morin, AB Sparks, V Korinek, N Barker, H Clevers, B Vogelstein, and KW Kinzler. Activation of beta-catenin-tcf signaling in colon cancer by mutations in beta-catenin or apc. *Science*, 275(5307):1787–1790, 1997.

B Morson. President's address. the polyp-cancer sequence in the large bowel. *Proc R Soc Med*, 67(6):451–457, 1974.

CA Moskaluk and SE Kern. Cancer gets mad: Dpc4 and other tgfbeta pathway genes in human cancer. *Biochim Biophys Acta*, 1288(3):M31–3, 1996.

A Moynihan, P Molloy, V Papangelis, Graeme Young, and Lawrence LaPointe. Upregulation of mesothelin, regiv, and transcobalamin in colon adenomas and cancer. In *Gastroenterology*. AGA, 2009. Poster: DDW (Chicago).

S Munemitsu, I Albert, B Souza, B Rubinfeld, and P Polakis. Regulation of intracellular beta-catenin levels by the adenomatous polyposis coli (apc) tumor-suppressor protein. *Proc Natl Acad Sci U S A*, 92(7):3046–3050, 1995.

S Muro, I Takemasa, S Oba, R Matoba, N Ueno, C Maruyama, R Yamashita, M Sekimoto, H Yamamoto, S Nakamori, M Monden, S Ishii, and K Kato. Identification of expressed genes linked to malignancy of human colorectal carcinoma by parametric clustering of quantitative expression data. *Genome Biol*, 4(3):R21, 2003.

T Muto, HJ Bussey, and BC Morson. The evolution of cancer of the colon and rectum. *Cancer*, 36(6):2251–2270, 1975.

Padmanabhan Nair, Sara Lagerholm, Sudhir Dutta, Samina Shami, Kirk Davis, Shuzhen Ma, and Mehran Malayeri. Coprocytobiology: on the nature of cellular elements from stools in the pathophysiology of colonic disease. *J Clin Gastroenterol*, 36(5 Suppl):S84–93; discussion S94–6, 2003.

George Nakos and David Joyner. *Linear Algebra With Applications*. Brooks/Cole Pub Co, 1998.

M Nannini, Maria A Pantaleo, Alessandra Maleddu, Annalisa Astolfi, Serna Formica, and Guido Biasco. Gene expression profiling in colorectal cancer using microarray technologies: Results and perspectives. *Cancer Treat Rev*, NIL(NIL):NIL, 2008.

BA Narayanan, NK Narayanan, B Simi, and BS Reddy. Modulation of inducible nitric oxide synthase and related proinflammatory genes by the omega-3 fatty acid docosahexaenoic acid in human colon cancer cells. *Cancer Res*, 63(5): 972–979, 2003.

E Neumann, S Lechner, IH Tarner, J Grifka, S Gay, J Ruschoff, B Renke, J Scholmerich, F Kullmann, and U Muller-Ladner. Evaluation of differentially expressed genes by a combination of cdna array and rap-pcr using the atlasimage 2.0 software. *J Autoimmun*, 21(2):161–166, 2003.

J Neyman and ES Pearson. On the problem of the most efficient tests of statisitcal hypotheses. *Philos. Trans. Roy. Soc. Series A*, 231:289–337., 1932.

NIH/NLM. Ncbi homepage. WWW, 2008. URL `http://www.ncbi.nlm.nih.gov/`.

TG Nishanian, JS Kim, A Foxworth, and T Waldman. Suppression of tumorigenesis and activation of wnt signaling by bone morphogenetic protein 4 in human cancer cells. *Cancer Biol Ther*, 3(7), 2004.

I Nishisho, Y Nakamura, Y Miyoshi, Y Miki, H Ando, A Horii, K Koyama, J Utsunomiya, S Baba, and P Hedge. Mutations of chromosome 5q21 genes in fap and colorectal cancer patients. *Science*, 253(5020):665–669, 1991.

Yasushi Nitanai, Yoshinori Satow, Hideki Adachi, and Masafumi Tsujimoto. Crystal structure of human renal dipeptidase involved in beta-lactam hydrolysis. *J Mol Biol*, 321(2):177–84, 2002.

DA Notterman, U Alon, AJ Sierk, and AJ Levine. Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays. *Cancer Res*, 61(7):3124–3130, 2001.

Roel Nusse. The wnt homepage. WWW, 2008. URL `http://www.stanford.edu/~rnusse/wntwindow.html`.

Shuji Ogino and Ajay Goel. Molecular classification and correlates in colorectal cancer. *J Mol Diagn*, 10(1):13–27, 2008.

Takahiro Ohmachi, Fumiaki Tanaka, Koshi Mimori, Hiroshi Inoue, Katsuhiko Yanaga, and Masaki Mori. Clinical significance of TROP2 expression in colorectal cancer. *Clin Cancer Res*, 12(10):3057–63, 2006.

Eiki Ojima, Yasuhiro Inoue, Chikao Miki, Masaki Mori, and Masato Kusunoki. Effectiveness of gene expression profiling for response prediction of rectal cancer to preoperative radiotherapy. *J Gastroenterol*, 42(9):730–6, 2007.

CF Ortega-Cava, S Ishihara, MA Rumi, K Kawashima, N Ishimura, H Kazumori, J Udagawa, Y Kadowaki, and Y Kinoshita. Strategic compartmentalization of toll-like receptor 4 in the mouse gut. *J Immunol*, 170(8):3977–3985, 2003.

H Oshima, M Oshima, M Kobayashi, M Tsutsumi, and MM Taketo. Morphological and molecular processes of polyp formation in apc(delta716) knockout mice. *Cancer Res*, 57(9):1644–1649, 1997.

M Oshima, H Oshima, K Kitagawa, M Kobayashi, C Itakura, and M Taketo. Loss of apc heterozygosity and abnormal tissue building in nascent intestinal polyps in mice carrying a truncated apc gene. *Proc Natl Acad Sci U S A*, 92 (10):4482–4486, 1995.

A Ougolkov, B Zhang, K Yamashita, V Bilim, M Mai, SY Fuchs, and T Minamoto. Associations among beta-trcp, an e3 ubiquitin ligase receptor, beta-catenin, and nf-kappab in colorectal cancer. *J Natl Cancer Inst*, 96(15):1161–1170, 2004.

YK Park, JL Franklin, SH Settle, SE Levy, E Chung, LH Jeyakumar, Y Shyr, MK Washington, RH Whitehead, BJ Aronow, and RJ Coffey. Gene expression profile analysis of mouse colon embryonic development. *Genesis*, 41(1):1–12, 2005.

R Parsons, GM Li, MJ Longley, WH Fang, N Papadopoulos, J Jen, A de la Chapelle, KW Kinzler, B Vogelstein, and P Modrich. Hypermutability and mismatch repair deficiency in rer+ tumor cells. *Cell*, 75(6):1227–1236, 1993.

Susanne Pedersen, Glenn Brown, Lloyd Graham, Robert Dunne, Peter Molloy, L Clark, Graeme P Young, and Lawrence LaPointe. A novel colorectal neoplasia gene (crng) with high sensitivity and specificity for both adenomas and cancers. In *Gastroenterology*. AGA, 2009a. Poster: DDW (Chicago).

Susanne Pedersen, Emma Richards, Aidan McEvoy, Robert Dunne, Glenn Brown, L Clark, Graeme P Young, and Lawrence LaPointe. Alternative splicing of s100a11 in colorectal adenomas and carcinomas. In *Gastroenterology*. AGA, 2009b. Poster: DDW (Chicago).

M Peifer. Developmental biology: colon construction. *Nature*, 420(6913):274–5, 277, 2002.

P Peltomaki. Deficient dna mismatch repair: a common etiologic factor for colon cancer. *Hum Mol Genet*, 10(7):735–740, 2001.

P Peltomaki. Role of dna mismatch repair defects in the pathogenesis of human cancer. *J Clin Oncol*, 21(6):1174–1179, 2003.

P Peltomaki and A de la Chapelle. Mutations predisposing to hereditary nonpolyposis colorectal cancer. *Adv Cancer Res*, 71:93–119, 1997.

P Peltomaki, LA Aaltonen, P Sistonen, L Pylkkanen, JP Mecklin, H Jarvinen, JS Green, JR Jass, JL Weber, FS Leach, and al. et. Genetic mapping of a locus predisposing to human colorectal cancer. *Science*, 260(5109):810–812, 1993.

D Pennica, TA Swanson, JW Welsh, MA Roy, DA Lawrence, J Lee, J Brush, LA Taneyhill, B Deuel, M Lew, C Watanabe, RL Cohen, MF Melhem, GG Finley, P Quirke, AD Goddard, KJ Hillan, AL Gurney, D Botstein, and AJ Levine. Wisp genes are members of the connective tissue growth factor family that are up-regulated in wnt-1-transformed cells and aberrantly expressed in human colon tumors. *Proc Natl Acad Sci U S A*, 95(25):14717–14722, 1998.

Margaret Sullivan Pepe, Ruth Etzioni, Zideng Feng, John D Potter, Mary Lou Thompson, Mark Thornquist, Marcy Winget, and Yutaka Yasui. Phases of biomarker development for early detection of cancer. *J Nat Can Inst*, 93(14): 1054–1061, 2001.

Perry J Pickhardt, Pamela A Nugent, Pauline A Mysliwiec, J Richard Choi, and William R Schindler. Location of adenomas missed by optical colonoscopy. *Ann Intern Med*, 141(5):352–9, 2004.

M Pierce, C Wang, M Stump, and A Kamb. Overexpression of the beta-catenin binding domain of cadherin selectively kills colorectal cancer cells. *Int J Cancer*, 107(2):229–237, 2003.

JA Pietenpol, RW Stein, E Moran, P Yaciuk, R Schlegel, RM Lyons, MR Pittelkow, K Munger, PM Howley, and HL Moses. Tgf-beta 1 inhibition of c-myc transcription and growth in keratinocytes is abrogated by viral transforming proteins with prb binding domains. *Cell*, 61(5):777–785, 1990.

M Plateroti, DC Rubin, I Duluc, R Singh, C Foltzer-Jourdainne, JN Freund, and M Kedinger. Subepithelial fibroblast cell lines from different levels of gut axis display regional characteristics. *Am J Physiol*, 274(5 Pt 1):G945–54, 1998.

J Platt. Sequential minimal optimization: A fast algorithm for training support vector machines. Technical Report MSR-TR-98-14, Microsoft Research, 1998.

J Platt. Fast training of support vector machines using sequential minimal optimization. In Scholkopf B, Burges C, and Smola A, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press, Cambridge, MA USA, 1999.

P Platzer, MB Upender, K Wilson, J Willis, J Lutterbaugh, A Nosrati, JK Willson, D Mack, T Ried, and S Markowitz. Silence of chromosomal amplifications in colon cancer. *Cancer Res*, 62(4):1134–1138, 2002.

P Polakis. Wnt signaling and cancer. *Genes Dev*, 14(15):1837–1851, 2000.

P Polakis. The many ways of wnt in cancer. *Curr Opin Genet Dev*, 17(1):45–51, Feb 2007. doi: 10.1016/j.gde.2006.12.007. URL `http://www.hubmed.org/display.cgi?uids=17208432`.

JD Potter. Colorectal cancer: molecules and populations. *J Natl Cancer Inst*, 91(11):916–932, 1999.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

Freddy Radtke and Hans Clevers. Self-renewal and cancer of the gut: Two sides of a coin. *Science*, 307(5717):1904–1909, 2005.

VM Rajendran, J Black, TA Ardito, P Sangan, SL Alper, C Schweinfest, M Kashgarian, and HJ Binder. Regulation of dra and ae1 in rat colon by dietary na depletion. *Am J Physiol Gastrointest Liver Physiol*, 279(5):G931–42, 2000.

S Ramaswamy, P Tamayo, R Rifkin, S Mukherjee, CH Yeang, M Angelo, C Ladd, M Reich, E Latulippe, JP Mesirov, T Poggio, W Gerald, M Loda, ES Lander, and TR Golub. Multiclass cancer diagnosis using tumor gene expression signatures. *Proc Natl Acad Sci U S A*, 98(26):15149–15154, 2001.

Keith N Rand, Thu Ho, Wenjia Qu, Susan M Mitchell, Rose White, Susan J Clark, and Peter L Molloy. Headloop suppression PCR and its application to selective amplification of methylated DNA sequences. *Nucleic Acids Res*, 33 (14):e127, 2005.

D F Ransohoff. Evaluating discovery-based research: when biologic reasoning cannot work. *Gastroenterology*, 127(4):1028–1028, Oct 2004a. URL `http://www.hubmed.org/display.cgi?uids=15480977`.

DF Ransohoff. Cancer. developing molecular biomarkers for cancer. *Science*, 299(5613):1679–1680, 2003.

DF Ransohoff. Rules of evidence for cancer molecular-marker discovery and validation. *Nat Rev Cancer*, 4(4):309–314, 2004b.

C. R. Rao. The utilization of multiple measurements in problems of biological classification (with discussion). *Journal of the Royal Statistical Society B*, 10: 159–203, 1948.

D Rasnick and PH Duesberg. How aneuploidy affects metabolic control and causes cancer. *Biochem J*, 340(Pt 3):621–630, 1999.

MA Reale and ER Fearon. Gene defects in colorectal tumorigenesis. In GP Young, P Rozen, and B Levin, editors, *Prevention and Early Detection of Colorectal Cancer*, pages 63–86. WB Saunders, 1997.

Tim Reichling, Kathleen Heppner Goss, Daniel J Carson, Robert W Holdcraft, Cathy Ley-Ebert, Dave Witte, Bruce J Aronow, and Joanna Groden. Transcriptional profiles of intestinal tumors in Apc(Min) mice are unique from those of embryonic intestine and identify novel gene targets dysregulated in human colorectal tumors. *Cancer Res*, 65(1):166–76, 2005.

DK Rex. Screening for colon cancer and evaluation of chemoprevention with coxibs. *J Pain Symptom Manage*, 23(4 Suppl):S41–50, 2002.

GJ Riggins, S Thiagalingam, E Rozenblum, CL Weinstein, SE Kern, SR Hamilton, JK Willson, SD Markowitz, KW Kinzler, and B Vogelstein. Mad-related genes in the human. *Nat Genet*, 13(3):347–349, 1996.

B D Ripley. *Pattern Recognition and Neural Networks*. Oxford Press, Cambridge, UK, 1st edition, 1996.

DJ Roberts. *Development of the Gastrointestinal Tract*. B C Decker, 1999.

CF Rochlitz, R Herrmann, and E de Kant. Overexpression and amplification of c-myc during progression of human colorectal cancer. *Oncology*, 53(6): 448–454, 1996.

J Roose and H Clevers. Tcf transcription factors: molecular switches in carcinogenesis. *Biochim Biophys Acta*, 1424(2-3):M23–37, 1999.

J Roose, G Huls, M van Beest, P Moerer, K van der Horn, R Goldschmeding, T Logtenberg, and H Clevers. Synergy between tumor suppressor apc and the beta-catenin-tcf4 target tcf1. *Science*, 285(5435):1923–1926, 1999.

F Rosenblatt. The perceptron: A probalistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958.

Sarah Ross and Caroline Hill. How the smads regulate transcription. *Int J Biochem Cell Bio*, 2007.

M A Rubin. Use of laser capture microdissection, cDNA microarrays, and tissue microarrays in advancing our understanding of prostate cancer. *J Pathol*, 195 (1):80–6, 2001.

J Sabates-Bellver, L G Van der Flier, M de Palo, E Cattaneo, C Maake, H Rehrauer, E Laczko, M A Kurowski, J M Bujnicki, M Menigatti, J Luz, T V Ranalli, V Gomes, A Pastorelli, R Faggiani, M Anti, J Jiricny, H Clevers, and G Marra. Transcriptome profile of human colorectal adenomas. *Mol Cancer Res*, 5(12):1263–1275, Dec 2007. doi: 10.1158/1541-7786.MCR-07-0267. URL `http://www.hubmed.org/display.cgi?uids=18171984`.

S Sachs. Screening for lung cancer: an old idea revisited. *Respir Care Clin N Am*, 9(1):27–50, 2003.

Sima Salahshor, Jason Goncalves, Runjan Chetty, Steven Gallinger, and James R Woodgett. Differential gene expression profile reveals deregulation of

pregnancy specific beta1 glycoprotein 9 early during colorectal carcinogenesis. *BMC Cancer*, 5(NIL):66, 2005.

I Salama, P S Malone, F Mihaimeed, and J L Jones. A review of the S100 proteins in cancer. *Eur J Surg Oncol*, 34(4):357–64, 2008.

WS Samowitz, K Curtin, HH Lin, MA Robertson, D Schaffer, M Nichols, K Gruenthal, MF Leppert, and ML Slattery. The colon cancer burden of genetically defined hereditary nonpolyposis colon cancer. *Gastroenterology*, 121(4):830–838, 2001.

Ian Saunders. A measure of the performance of biomarkers for disease. *Cancer Biomarkers*, 2:145–150, 2006.

Ian Saunders. Bayes estimation of d value. Personal communication, provided R code., 2008.

L Schiff, RJ Stevens, N Shapiro, and S Goodman. Observations on the oral administration of citrated blood in man. *Am J of Med Sci*, 203:409–412, 1942.

Oliver Schoor, Toni Weinschenk, Jorg Hennenlotter, Stefan Corvin, Arnulf Stenzl, Hans-Georg Rammensee, and Stefan Stevanovic. Moderate degradation does not preclude microarray analysis of small amounts of RNA. *Biotechniques*, 35(6):1192–6, 1198–201, 2003.

Stefania Segditsas, Oliver Sieber, Maesha Deheragoda, Phil East, Andrew Rowan, Rosemary Jeffery, Emma Nye, Susan Clark, Bradley Spencer-Dene, Gordon Stamp, Richard Poulsom, Nirosha Suraweera, Andrew Silver, Mohammad Ilyas, and Ian Tomlinson. Putative direct and indirect Wnt targets identified through consistent gene expression changes in APC-mutant intestinal adenomas from humans and mice. *Hum Mol Genet*, 17(24):3864–75, 2008.

I M Seiden-Long, K R Brown, W Shih, D A Wigle, N Radulovich, I Jurisica, and M-S Tsao. Transcriptional targets of hepatocyte growth factor signaling and ki-ras oncogene activation in colorectal cancer. *Oncogene*, 25(1):91–102, 2006. ISSN 0950-9232 (Print). doi: 10.1038/sj.onc.1209005.

Jinesh N Shah, Genze Shao, Tom K Hei, and Yongliang Zhao. Methylation screening of the TGFBI promoter in human lung and prostate cancer by methylation-specific PCR. *BMC Cancer*, 8(NIL):284, 2008.

Sumana Shashidhar, Gustavo Lorente, Usha Nagavarapu, April Nelson, Jane Kuo, Jeramiah Cummins, Karoly Nikolich, Roman Urfer, and Erik D Foehr. GPR56 is a GPCR that is overexpressed in gliomas and functions in tumor cell adhesion. *Oncogene*, 24(10):1673–82, 2005.

L Shi, L H Reid, W D Jones, R Shippy, J A Warrington, S C Baker, P J Collins, F de Longueville, E S Kawasaki, K Y Lee, Y Luo, Y A Sun, J C Willey, R A Setterquist, G M Fischer, W Tong, Y P Dragan, D J Dix, F W Frueh, F M Goodsaid, D Herman, R V Jensen, C D Johnson, E K Lobenhofer, R K Puri, U Schrf, J Thierry-Mieg, C Wang, M Wilson, P K Wolber, L Zhang, S Amur, W Bao, C C Barbacioru, A B Lucas, V Bertholet, C Boysen, B Bromley, D Brown, A Brunner, R Canales, X M Cao, T A Cebula, J J Chen, J Cheng, T M Chu, E Chudin, J Corson, J C Corton, L J Croner, C Davies, T S Davison, G Delenstarr, X Deng, D Dorris, A C Eklund, X H Fan, H Fang, S Fulmer-Smentek, J C Fuscoe, K Gallagher, W Ge, L Guo, X Guo, J Hager, P K Haje, J Han, T Han, H C Harbottle, S C Harris, E Hatchwell, C A Hauser, S Hester, H Hong, P Hurban, S A Jackson, H Ji, C R Knight, W P Kuo, J E LeClerc, S Levy, Q Z Li, C Liu, Y Liu, M J Lombardi, Y Ma, S R Magnuson, B Maqsodi, T McDaniel, N Mei, O Myklebost, B Ning, N Novoradovskaya, M S Orr, T W Osborn, A Papallo, T A Patterson, R G Perkins, E H Peters, R Peterson, K L Philips, P S Pine, L Pusztai, F Qian, H Ren, M Rosen, B A Rosenzweig, R R Samaha, M Schena, G P Schroth, S Shchegrova, D D Smith, F Staedtler, Z Su, H Sun, Z Szallasi, Z Tezak, D Thierry-Mieg, K L Thompson, I Tikhonova, Y Turpaz, B Vallanat, C Van, S J Walker, S J Wang, Y Wang, R Wolfinger, A Wong, J Wu, C Xiao, Q Xie, J Xu, W Yang, L Zhang, S Zhong, Y Zong, and W Slikker. The microarray quality control (maqc) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol*, 24(9):1151–1161, Sep 2006. doi: 10.1038/nbt1239. URL `http://www.hubmed.org/display.cgi?uids=16964229`.

M Shtutman, J Zhurinsky, I Simcha, C Albanese, M D'Amico, R Pestell, and A Ben-Ze'ev. The cyclin d1 gene is a target of the beta-catenin/lef-1 pathway. *Proc Natl Acad Sci U S A*, 96(10):5522–5527, 1999.

DG Silberg, GP Swain, ER Suh, and PG Traber. Cdx1 and cdx2 expression during intestinal development. *Gastroenterology*, 119(4):961–971, 2000.

R Simon, MD Radmacher, K Dobbin, and LM McShane. Pitfalls in the use of dna microarray data for diagnostic and prognostic classification. *J Natl Cancer Inst*, 95(1):14–18, 2003.

S Singh, R Poulsom, AM Hanby, LA Rogers, NA Wright, MC Sheppard, and MJ Langman. Expression of oestrogen receptor and oestrogen-inducible genes ps2 and erd5 in large bowel mucosa and cancer. *J Pathol*, 184(2):153–160, 1998.

Natalia A Skrypina, Angelica V Timofeeva, George L Khaspekov, Larissa P Savochkina, and Robert Sh Beabealashvilli. Total RNA suitable for molecular biology analysis. *J Biotechnol*, 105(1-2):1–9, 2003.

MJ Smalley and TC Dale. Wnt signalling in mammalian development and cancer. *Cancer Metastasis Rev*, 18(2):215–230, 1999.

Alicia Smith, Graeme P Young, Stephen R Cole, and Peter Bampton. Comparison of a brush-sampling fecal immunochemical test for hemoglobin with a sensitive guaiac-based fecal occult blood test in detection of colorectal neoplasia. *Cancer*, 107(9):2152–9, 2006.

G K Smyth. Linear models and empirical bayes methods for assessing differential epxressionin microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1), 2004.

G K Smyth. Limma: linear models for microarray data. In R Gentleman, V Carey, S Dudoit, R Irizarry, and W Huber, editors, *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, pages 397–420. Springer, New York, 2005.

K Soreide, E A M Janssen, H Soiland, H Korner, and J P A Baak. Microsatellite instability in colorectal cancer. *Br J Surg*, 93(4):395–406, 2006.

K Soreide, BS Nedrebo, Jens-Christian Knapp, T Glomsaker, Jon Arne Soreide, and Hartwig Korner. Evolving molecular classification by genomic and proteomic biomarkers in colorectal cancer: Potential implications for the surgical oncologist. *Surg Oncol*, EPUB(NIL):NIL, 2008.

C Sotiriou and M J Piccart. Taking gene-expression profiling to the clinic: when will molecular signatures become relevant to patient care? *Nat Rev Cancer*, 7 (7):545–553, Jul 2007. doi: 10.1038/nrc2173. URL `http://www.hubmed.org/display.cgi?uids=17585334`.

RF Souza, J Yin, KN Smolinski, TT Zou, S Wang, YQ Shi, MG Rhyu, J Cottrell, JM Abraham, K Biden, L Simms, B Leggett, GS Bova, T Frank, SM Powell, H Sugimura, J Young, N Harpaz, K Shimizu, N Matsubara, and SJ Meltzer. Frequent mutation of the e2f-4 cell cycle gene in primary human gastrointestinal tumors. *Cancer Res*, 57(12):2350–2353, 1997.

Bonnie Spring. Health Decision Making: Lynchpin of Evidence-Based Practice. *Med Decis Making*, 28(6):866–874, 2008. doi: 10.1177/0272989X08326146. URL `http://mdm.sagepub.com/cgi/content/abstract/28/6/866`.

J E Stajich, D Block, K Boulez, S E Brenner, S A Chervitz, C Dagdigian, G Fuellen, J G Gilbert, I Korf, H Lapp, H Lehväslaiho, C Matsalla, C J Mungall, B I Osborne, M R Pocock, P Schattner, M Senger, L D Stein, E Stupka, M D Wilkinson, and E Birney. The bioperl toolkit: Perl modules for the life sciences. *Genome Res*, 12(10):1611–1618, Oct 2002. doi: 10.1101/gr.361602. URL `http://www.hubmed.org/display.cgi?uids=12368254`.

AI Su, JB Welsh, LM Sapinoso, SG Kern, P Dimitrov, H Lapp, PG Schultz, SM Powell, CA Moskaluk, HF Jr Frierson, and GM Hampton. Molecular classification of human carcinomas by use of gene expression signatures. *Cancer Res*, 61(20):7388–7393, 2001.

LK Su, B Vogelstein, and KW Kinzler. Association of the apc tumor suppressor protein with catenins. *Science*, 262(5140):1734–1737, 1993.

Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, 2005. doi: 10.1073/pnas.0506580102. URL `http://www.pnas.org/content/102/43/15545.abstract`.

Neetu Sud, Rinu Sharma, Riju Ray, Tushar Kant Chattopadhyay, and Ranju Ralhan. Differential expression of G-protein coupled receptor 56 in human esophageal squamous cell carcinoma. *Cancer Lett*, 233(2):265–70, 2006.

Takeyuki Sugiura, Aya Yamaguchi, and Kentaro Miyamoto. A cancer-associated RING finger protein, RNF43, is a ubiquitin ligase that interacts with a nuclear protein, HAP95. *Exp Cell Res*, 314(7):1519–28, 2008.

Yuko Sugiyama, Buckminster Farrow, Carlos Murillo, Jing Li, Hiroaki Watanabe, Kazuo Sugiyama, and B Mark Evers. Analysis of differential gene expression patterns in colon cancer and cancer stroma using microdissected tissues. *Gastroenterology*, 128(2):480–6, 2005.

J Taipale and P A Beachy. The Hedgehog and Wnt signalling pathways in cancer. *Nature*, 411(6835):349–54, 2001.

I Takemasa, H Higuchi, H Yamamoto, M Sekimoto, N Tomita, S Nakamori, R Matoba, M Monden, and K Matsubara. Construction of preferential cdna microarray specialized for human colorectal carcinoma: molecular sketch of colorectal cancer. *Biochem Biophys Res Commun*, 285(5):1244–1249, 2001.

M Tanaka, K Adzuma, M Iwami, K Yoshimoto, Y Monden, and M Itakura. Human calgizzarin; one colorectal cancer-related gene selected by a large scale random cDNA sequencing and northern blot analysis. *Cancer Lett*, 89(2):195–200, 1995.

O Tetsu and F McCormick. Beta-catenin regulates expression of cyclin d1 in colon carcinoma cells. *Nature*, 398(6726):422–426, 1999.

SN Thibodeau, G Bren, and D Schaid. Microsatellite instability in cancer of the proximal colon. *Science*, 260(5109):816–819, 1993.

R Tibshirani. Regression shrinkage and selection via the lasso. *J Royal Statistical Society. Series B (Methodological)*, 58(1):267–268, 1996.

A V Tinker, A Boussioutas, and D D Bowtell. The challenges of gene expression microarrays for the study of human cancer. *Cancer Cell*, 9(5):333–339, May 2006. doi: 10.1016/j.ccr.2006.05.001. URL `http://www.hubmed.org/display.cgi?uids=16697954`.

M Toyota, N Ahuja, M Ohe-Toyota, JG Herman, SB Baylin, and JP Issa. Cpg island methylator phenotype in colorectal cancer. *Proc Natl Acad Sci U S A*, 96(15):8681–8686, 1999.

PG Traber. Transcriptional regulation in intestinal development. implications for colorectal cancer. *Adv Exp Med Biol*, 470:1–14, 1999.

PC Turner, AG McLennan, AD Bates, and MRH White. *Molecular biology*. Springer-Verlag (BIOS Scientific Publishers), School of Biological Sciences, University of Liverpool, Liverpool, UK., 2nd edition, 2000.

Olga Turovskaya, Dirk Foell, Pratima Sinha, Thomas Vogl, Robbin Newlin, Jonamani Nayak, Mien Nguyen, Anna Olsson, Peter P Nawroth, Angelika Bierhaus, Nissi Varki, Mitchell Kronenberg, Hudson H Freeze, and Geetha Srikrishna. RAGE, carboxylated glycans and S100A8/A9 play essential roles in colitis-associated carcinogenesis. *Carcinogenesis*, 29(10):2035–43, 2008.

M van de Wetering, E Sancho, C Verweij, W de Lau, I Oving, A Hurlstone, K van der Horn, E Batlle, D Coudreuse, AP Haramis, M Tjon-Pon-Fong, P Moerer, M van den Born, G Soete, S Pals, M Eilers, R Medema, and H Clevers. The beta-catenin/tcf-4 complex imposes a crypt progenitor phenotype on colorectal cancer cells. *Cell*, 111(2):241–250, 2002.

I M M van Leeuwen, H M Byrne, O E Jensen, and J R King. Crypt dynamics and colorectal cancer: advances in mathematical modelling. *Cell Prolif*, 39 (3):157–81, 2006.

V Vapnik. *The Nature of Statistical Learning Theory*, pages 136–7. Statistics for Engineering and Information Science. Springer-Verlag, New York, 2nd edition, 1995.

VE Velculescu, L Zhang, B Vogelstein, and KW Kinzler. Serial analysis of gene expression. *Science*, 270(5235):484–487, 1995.

W N Venables and B D Ripley. *Modern Applied Statistics with S*, chapter 11. Statistics and Computing. Springer-Verlag, New York, fourth edition, 2002.

B Vogelstein, ER Fearon, SR Hamilton, SE Kern, AC Preisinger, M Leppert, Y Nakamura, R White, AM Smits, and JL Bos. Genetic alterations during colorectal-tumor development. *N Engl J Med*, 319(9):525–532, 1988.

L Waltzer and M Bienz. The control of beta-catenin and tcf during embryonic development and cancer. *Cancer Metastasis Rev*, 18(2)(2):231–246, 1999.

Hans H Wandall, Sally Dabelsteen, Jens Ahm Sorensen, Annelise Krogdahl, Ulla Mandel, and Erik Dabelsteen. Molecular basis for the presence of glycosylated onco-foetal fibronectin in oral carcinomas: the production of glycosylated onco-foetal fibronectin by carcinoma cells. *Oral Oncol*, 43(3):301–9, 2007.

Y Wang, T Jatkoe, Y Zhang, MG Mutch, D Talantov, J Jiang, HL McLeod, and D Atkins. Gene expression profiles and molecular markers to predict recurrence of dukes' b colon cancer. *J Clin Oncol*, 22(9):1564–1571, 2004.

BJ Warner, SW Blain, J Seoane, and J Massague. Myc downregulation by transforming growth factor beta required for activation of the p15(ink4b) g(1) arrest pathway. *Mol Cell Biol*, 19(9):5913–5922, 1999.

T Watanabe, T Kobunai, E Toda, T Kanazawa, Y Kazama, J Tanaka, T Tanaka, Y Yamamoto, K Hata, T Kojima, T Yokoyama, T Konishi, Y Okayama,

Y Sugimoto, T Oka, S Sasaki, Y Ajioka, T Muto, and H Nagawa. Gene expression signature and the prediction of ulcerative colitis-associated colorectal cancer by dna microarray. *Clin Cancer Res*, 13(2 Pt 1):415–420, 2007.

ML Waterman. Lymphoid enhancer factor/t cell factor expression in colorectal cancer. *Cancer Metastasis Rev*, 23(1-2):41–52, 2004.

JN Weinstein. Searching for pharmacogenomic markers: the synergy between omic and hypothesis-driven research. *Dis Markers*, 17(2):77–88, 2001.

Daniel J Weisenberger, Kimberly D Siegmund, Mihaela Campan, Joanne Young, Tiffany I Long, Mark A Faasse, Gyeong Hoon Kang, Martin Widschwendter, Deborah Weener, Daniel Buchanan, Hoey Koh, Lisa Simms, Melissa Barker, Barbara Leggett, Joan Levine, Myungjin Kim, Amy J French, Stephen N Thibodeau, Jeremy Jass, Robert Haile, and Peter W Laird. CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer. *Nat Genet*, 38(7):787–93, 2006.

Evelyn P Whitlock, Jennifer S Lin, Elizabeth Liles, Tracy L Beil, and Rongwei Fu. Screening for colorectal cancer: a targeted, updated systematic review for the U.S. Preventive Services Task Force. *Ann Intern Med*, 149(9):638–58, 2008.

VJ Wielenga, R Smits, V Korinek, L Smit, M Kielman, R Fodde, H Clevers, and ST Pals. Expression of cd44 in apc and tcf mutant mice implies regulation by the wnt pathway. *Am J Pathol*, 154(2):515–523, 1999.

Anja H Wiese, Johannes Auer, Silke Lassmann, Jorg Nahrig, Robert Rosenberg, Heinz Hofler, Rudiger Ruger, and Martin Werner. Identification of gene signatures for invasive colorectal tumor cells. *Cancer Detect Prev*, 31(4):282–95, 2007.

NS Williams, RB Gaynor, S Scoggin, U Verma, T Gokaslan, C Simmang, J Fleming, D Tavana, E Frenkel, and C Becerra. Identification and validation of genes

involved in the pathogenesis of colorectal cancer using cdna microarrays and rna interference. *Clin Cancer Res*, 9(3):931–946, 2003.

SJ Williams, MA McGuckin, DC Gotley, HJ Eyre, GR Sutherland, and TM Antalis. Two novel mucin genes down-regulated in colorectal cancer identified by differential display. *Cancer Res*, 59(16):4083–4089, 1999.

Claire L Wilson and Crispin J Miller. Simpleaffy: a BioConductor package for Affymetrix Quality Control and data analysis. *Bioinformatics*, 21(18):3683–5, 2005.

Sidney J. Winawer, Ann G. Zauber, Robert H. Fletcher, Jonathon S. Stillman, Michael J. O'Brien, Bernard Levin, Robert A. Smith, David A. Lieberman, Randall W. Burt, Theodore R. Levin, John H. Bond, Durado Brooks, Tim Byers, Neil Hyman, Lynne Kirk, Alan Thorson, Clifford Simmang, David Johnson, and Douglas K. Rex. Guidelines for Colonoscopy Surveillance after Polypectomy: A Consensus Update by the US Multi-Society Task Force on Colorectal Cancer and the American Cancer Society,. *CA Cancer J Clin*, 56 (3):143–159, 2006. doi: 10.3322/canjclin.56.3.143. URL `http://caonline.amcancersoc.org/cgi/content/abstract/56/3/143`.

SM Woerner, A Benner, C Sutter, M Schiller, YP Yuan, G Keller, P Bork, MK Doeberitz, and JF Gebert. Pathogenesis of dna repair-deficient cancers: a statistical meta-analysis of putative real common target genes. *Oncogene*, 22(15):2226–2235, 2003.

NA Wong and M Pignatelli. Beta-catenin–a linchpin in colorectal carcinogenesis? *Am J Pathol*, 160(2):389–401, 2002.

L Xu, RB Corcoran, JW Welsh, D Pennica, and AJ Levine. Wisp-1 is a wnt-1- and beta-catenin-responsive oncogene. *Genes Dev*, 14(5):585–595, 2000.

Lei Xu and Richard O Hynes. GPR56 and TG2: possible roles in suppression of tumor growth by the microenvironment. *Cell Cycle*, 6(2):160–5, 2007.

Ryuichiro Yagyu, Yoichi Furukawa, Yu-Min Lin, Takashi Shimokawa, Takehira

Yamamura, and Yusuke Nakamura. A novel oncoprotein RNF43 functions in an autocrine manner in colorectal cancer. *Int J Oncol*, 25(5):1343–8, 2004.

Satoshi Yajima, Mie Ishii, Hisayuki Matsushita, Kazuhiko Aoyagi, Kazuhiko Yoshimatsu, Hironori Kaneko, Nobuko Yamamoto, Tatsuo Teramoto, Teruhiko Yoshida, Yasuhiro Matsumura, and Hiroki Sasaki. Expression profiling of fecal colonocytes for RNA-based screening of colorectal cancer. *Int J Oncol*, 31(5):1029–37, 2007.

Tadataka Yamada, David Alpers, Neil Kaplowitz, Loren Laine, Chung Owyang, and Down W Powell. *Textbook of Gastroenterology.* Lipincott Wiliams and Wilkins, Philadelphia, 4th edition, 2003.

D Yan, M Wiesmann, M Rohan, V Chan, AB Jefferson, L Guo, D Sakamoto, RH Caothien, JH Fuller, C Reinhard, PD Garcia, FM Randazzo, J Escobedo, WJ Fantl, and LT Williams. Elevated expression of axin2 and hnkd mrna provides evidence that wnt/beta -catenin signaling is activated in human colon tumors. *Proc Natl Acad Sci U S A*, 98(26):14973–14978, 2001.

G Yang, Y Xu, X Chen, and G Hu. IFITM1 plays an essential role in the antiproliferative action of interferon-gamma. *Oncogene*, 26(4):594–603, 2007.

Ivana V Yang, Emily Chen, Jeremy P Hasseman, Wei Liang, Bryan C Frank, Shuibang Wang, Vasily Sharov, Alexander I Saeed, Joseph White, Jerry Li, Norman H Lee, Timothy J Yeatman, and John Quackenbush. Within the fold: assessing differential expression measures and reproducibility in microarray assays. *Genome Biol*, 3(11):research0062, 2002.

W Yang, A Velcich, J Mariadason, C Nicholas, G Corner, M Houston, W Edelmann, R Kucherlapati, PR Holt, and LH Augenlicht. p21(waf1/cip1) is an important determinant of intestinal cell response to sulindac in vitro and in vivo. *Cancer Res*, 61(16):6297–6302, 2001.

TJ Yeatman and W Mao. Identification fo a differentially-expressed message associated with colon cancer liver metastasis using an improved method of differential display. *Nucleic Acids Res*, 23(19):4007–8, 1995.

GP Young, P Rozen, and B Levin. *Prevention and Early Detection of Colorectal Cancer*. WB Saunders, New York, 1997.

Kun Yu, Kumaresan Ganesan, Lay Keng Tan, Mirtha Laban, Jeanie Wu, Xiao Dong Zhao, Hongmin Li, Carol Ho Wing Leung, Yansong Zhu, Chia Lin Wei, Shing Chuan Hooi, Lance Miller, and Patrick Tan. A precisely regulated gene expression cassette potently modulates metastasis and survival in multiple solid cancers. *PLoS Genet*, 4(7):e1000129, 2008.

Lin Zhang, Wei Zhou, Victor E Velculescu, Scott E Kern, Ralph H Hruban, Stanley R Hamilton, Bert Vogelstein, and Kenneth W. Kinzler. Gene expression profiles in normal and cancer cells. *Science*, 276(1268):1268–1272, 1997.

T Zhang, T Otevrel, Z Gao, Z Gao, SM Ehrlich, JZ Fields, and BM Boman. Evidence that apc regulates survivin expression: a possible mechanism contributing to the stem cell origin of colon cancer. *Cancer Res*, 61(24):8664–8667, 2001.

TT Zou, FM Selaru, Y Xu, V Shustova, J Yin, Y Mori, D Shibata, F Sato, S Wang, A Olaru, E Deacu, TC Liu, JM Abraham, and SJ Meltzer. Application of cdna microarrays to generate a molecular taxonomy capable of distinguishing between colon cancer and normal colon. *Oncogene*, 21(31): 4855–4862, 2002.

J Zumbrunn, K Kinoshita, AA Hyman, and IS Nathke. Binding of the adenomatous polyposis coli protein to microtubules increases microtubule stability and is regulated by gsk3 beta phosphorylation. *Curr Biol*, 11(1):44–49, 2001.