Post-GWAS functional characterisation of colorectal cancer risk loci

Li Yin Ooi

Thesis submitted for the degree of

Doctor of Philosophy

School of Medicine and Veterinary Medicine

The University of Edinburgh

2016

Dedication

I dedicate this thesis to Euan, for giving mummy a good run of nights, which made the writing of this thesis possible.

Declaration

I declare that this thesis was composed entirely by myself and that the research presented is my own unless otherwise stated. No part of this research has been submitted for any other degree or professional qualification.

Li Yin Ooi

Jan 2016

The following publications are derived from the research presented in this thesis:

Dunlop MG, Dobbins SE, Farrington SM, Jones AM, Palles C, Whiffin N, Tenesa A, Spain S, Broderick P, **Ooi LY**, Domingo E, Smillie C, Henrion M, Frampton M, Martin L, Grimes G, Gorman M, Semple C, Ma YP, Barclay E, Prendergast J, Cazier JB, Olver B, Penegar S, Lubbe S, Chander I, Carvajal-Carmona LG, Ballereau S, Lloyd A, Vijayakrishnan J, Zgaga L, Rudan I, Theodoratou E; Colorectal Tumour Gene Identification (CORGI) Consortium, Starr JM, Deary I, Kirac I, Kovacević D, Aaltonen LA, Renkonen-Sinisalo L, Mecklin JP, Matsuda K, Nakamura Y, Okada Y, Gallinger S, Duggan DJ, Conti D, Newcomb P, Hopper J, Jenkins MA, Schumacher F, Casey G, Easton D, Shah M, Pharoah P, Lindblom A, Liu T; Swedish Low-Risk Colorectal Cancer Study Group, Smith CG, West H, Cheadle JP; COIN Collaborative Group, Midgley R, Kerr DJ, Campbell H, Tomlinson IP, Houlston RS. Common variation near CDKN1A, POLD3 and SHROOM2 influences colorectal cancer risk. *Nat Genet* 2012; 44(7): 770-6.

Zgaga L, Theodoratou E, Farrington SM, Din FV, **Ooi LY**, Glodzik D, Johnston S, Tenesa A, Campbell H, Dunlop MG. Plasma vitamin D concentration influences survival outcome after a diagnosis of colorectal cancer. *J Clin Oncol* 2014; 32(23): 2430-9.

Timofeeva MN, Kinnersley B, Farrington SM, Whiffin N, Palles C, Svinti V, Lloyd A, Gorman M, **Ooi LY**, Hosking F, Barclay E, Zgaga L, Dobbins S, Martin L, Theodoratou E, Broderick P, Tenesa A, Smillie C, Grimes G, Hayward C, Campbell A, Porteous D, Deary IJ, Harris SE, Northwood EL, Barrett JH, Smith G, Wolf R, Forman D, Morreau H, Ruano D, Tops C, Wijnen J, Schrumpf M, Boot A, Vasen HF, Hes FJ, van Wezel T, Franke A, Lieb W, Schafmayer C, Hampe J, Buch S, Propping P, Hemminki K, Försti A, Westers H, Hofstra R, Pinheiro M, Pinto C, Teixeira M, Ruiz-Ponte C, Fernández-Rozadilla C, Carracedo A, Castells A, Castellví-Bel S, Campbell H, Bishop DT, Tomlinson IP, Dunlop MG, Houlston RS. Recurrent Coding Sequence Variation Explains Only A Small Fraction of the Genetic Architecture of Colorectal Cancer. *Sci Rep* 2015; 5: 16286.

ABSTRACT

Large bowel cancer, or colorectal cancer (CRC) is the third most common cause of cancer worldwide and the fourth biggest cause of cancer mortality. Twin studies have shown that the heritable contribution is ~35%, with ~5% of cases due to rare, high-penetrance mutations. In the last decade, the use of genome-wide association studies on large, well-characterised case-control cohorts of CRC has facilitated the identification of over 25 common genetic variants that carry with them an increased predisposition to colorectal cancer, invoking the common-disease common variant paradigm. As almost all of these variants lie within non-coding regions, the underlying causal mechanism is to-date poorly understood for the majority of these loci, and it is thought that they mediate risk by influencing gene expression levels.

To test this hypothesis, an agnostic approach that utilises expression quantitative trait loci (eQTL) analysis was first carried on 115 normal colorectal mucosa samples and 59 peripheral blood mononuclear cells (PBMC). As these heritable variation on gene expression are likely to be subtle, there is a strong emphasis on the technical methodology to minimise experimentally-induced non-biological variations, including the extraction of high-quality RNA from primary tissue, the selection and validation of reference genes for normalisation of gene expression quantification, as well as internal validation of the samples and data processing. Thereafter, the association between the 25 CRC risk variants and the expression of their *cis*-genes were examined systematically, demonstrating that ten of these variants are also tissue-specific eQTLs. This intermediate phenotype strongly suggests that they confer risk, at least in part, by modifying regulatory mechanisms. One of the best eQTL associations (Xp22.2) is investigated in further detail to reveal a novel indel polymorphism (Indel24) at the distal promoter region of target gene *SHROOM2* that influenced both transcript abundance and CRC risk more than the original tagging SNP. Functional verification with gene reporter assays indicated that Indel24 displays differential allelic control over transcriptional activity. Further *in silico* analysis and mutations to the reporter gene constructs provided evidence that Indel24 modulates transcription by modifying the spacing between CCAAT motifs and the consequent binding affinity of NF-Y transcription factor. siRNA depletion of NF-Y was associated with a reduction in transcriptional activity of the Indel24 gene construct as well as endogenous *SHROOM2*, which is strongly supportive of the interaction between Indel24 and NF-Y in the transcriptional activation of *SHROOM2*. Preliminary evidence is suggestive of *SHROOM2* being expressed at the top of the intestinal epithelial crypt and playing a role in cell cycle regulation.

Hypothesis-driven approaches can also be of utility in demonstrating functionality of CRC risk variants, complementing the hypothesis-free approach of eQTL analysis. Guided by a recently discovered gene-environment interaction between the 16q22.1 risk variant and circulating vitamin D levels, the influence of the rs9929218 SNP on *CDH1* gene expression was examined, in relation to the expression of putative regulatory genes derived from *in silico* analysis and studies of other target genes. Although there was no direct association between rs9929218 and *CDH1* expression, there were multiple two-way interactions that were together suggestive of rs9929218 influencing the *VDR/FOXO4* regulation of *CDH1*. This provides functional support for the mechanism underlying the epidemiological observation of the gene-environment interaction between 16q22.1 and vitamin D, and demonstrates a candidate-based approach in deciphering the link between genetic locus and CRC susceptibility.

In summary, the research presented in this thesis has validated the experimental rationale of utilising expression studies of normal colorectal mucosa to hone in on the molecular mechanisms and susceptibility genes underlying the association between common genetic variation and CRC risk.

LAY ABSTRACT

Large bowel cancer is the third most common cause of cancer worldwide and the fourth biggest cause of cancer deaths. Large-scale comparative studies of people with and without colorectal cancer have shown that there are inherited genetic factors that predispose one to the disease. These genetic factors are present at varying frequencies in the general population with varying effects on disease risk; rare genetic mutations have a big impact on the lifetime chances of developing the disease, whereas common normal DNA sequence differences have a smaller influence on disease susceptibility.

Although the risk conferred by these common DNA differences are individually modest, collectively they have a significant influence on the risk of developing the disease. How these variants lead to the development of large bowel cancer is poorly understood, and this study seeks to shed light on the underlying mechanisms. Understanding how these heritable factors lead to disease is important as not only will it improve our understanding of how cancer develops, it will also inform the design of preventative and therapeutic strategies.

By analysing the cells of the human large bowel and blood, this study demonstrates that some of these common genetic differences linked to large bowel cancer do not alter the *function* of genes, but instead influence the *levels* of gene products that are expressed. Further investigation of one of the genetic variants with the strongest influence on gene expression identifies the underlying molecular mechanism and the gene it influences (known as SHROOM2). The research presented in this thesis also presents a framework of investigation into the function of this gene in the large bowel, and how differences in its expression could lead to large bowel cancer. Finally, this thesis describes the investigation of the molecular mechanism underlying the synergistic effect of DNA variation and vitamin D levels on the risk of developing large bowel cancer. This is an important aspect to address as it is known that large bowel cancer arises from a combination of genetic and environmental factors, and a clearer understanding of this complex relationship will ultimately be of public health benefit.

ACKNOWLEDGEMENTS

First and foremost, I cannot be more thankful to my supervisors Dr Susan Farrington and Professor Malcolm Dunlop for their dedicated support during the course of this PhD. You have both been inspiring in different yet complementing ways. Thank you for teaching me the principles of good research, for offering kind encouragement, for critically analysing my work, for protecting my time, and for acknowledging my efforts. Special thanks to Susan for tirelessly proofreading my thesis! A round of applause must also go to Miss Farhat Din, for her selfless encouragement and advice throughout, and for helping with the organoid set-up.

I would like to give my sincerest thanks to all the patients who have kindly donated tissue for my research, without which none of this research would be possible. The Western General surgical, pathology, Tissue Governance and ECMC teams have facilitated the clinical aspects of sample collection and for that I am very grateful.

I would also like to express my gratitude to Marion Walker, Stuart Reid, Asta Valenciute and Maria Timofeeva for their continued help, technical support as well as friendship. Many thanks goes to Marion for sorting out my samples and putting up with my last-minute orders, as well as always laughing at my jokes and my life's little disasters. Particular thanks to Stuart and his brilliant cloning green fingers, his help has facilitated a lot of my research and the time working with him has been a pleasure. I have learnt the importance of being robust (in experimental technique and life in general) from Asta, thank you for always sharing antibodies as well as having me over at dinner parties! Special thanks go to Maria for helping me with statistical genetics, and for the company during the period of writing-up.

Many thanks to everyone else in the lab and the wider unit that have assisted with various aspects of this work, and who have endured my pestering without complaint.

An extra special thank you to my family –Dad, Mum, Su Yin and Wei Yin for your unconditional love, support and encouragement. Last but not least, thank you to Siang Ling who has given me unwavering support through all these years. Your belief in me has made me who I am today.

ABBREVIATIONS

25-OHD	25-hydroxyvitamin D
3C	Chromosome conformation capture
5-FU	Fluorouracil
Adj	Adjusted
ANOVA	Analysis of variance
APC	Adenomatous polyposis coli
APS	Ammonium persulphate
aRNA	Amplified RNA
AvgExp	Average Expression
bp	Base pair
BSA	Bovine serum albumin
CAC	Colitis-associated cancer
cDNA	Complementary DNA
ChIP	Chromatin-immunoprecipitation
ChIP-seq	Chromatin immunoprecipitation and massively parallel sequencing
Chr	Chromosome
CHRPE	Congenital Hypertrophy of Retinal Pigment Epithelium
CIMP	CpG island methylator phenotype
CIN	Chromosomal instability
COSHH	Control of substances hazardous to health
COX	Cyclooxygenase
CpG	—C—phosphate—G—
CRC	Colorectal cancer
CRISPR	Clustered regularly interspaced short palindromic repeats
CV	Coefficient of variation
DAVID	Database for annotation, visualization and integrated discovery
DMSO	Dimethylsulfoxide
DNA	Deoxyribonucleic acid

DNase	Deoxyribonuclease
dNTP	Deoxynucleotide Triphosphate
EBV	Ebstein-Barr virus
EDTA	Ethylenediaminetetraacetic acid
EMSA	Electrophoretic mobility shift assay
ENCODE	Encyclopedia of DNA Elements
eQTL	Expression quantitative trait loci
ETOH	Ethanol
Exo-SAP	Exonuclease I - Shrimp Alkaline Phosphatase
FAP	Familial Adenomatous Polyposis
FC	Fold change
FDR	False discovery rate
FUCCI	Fluorescence Ubiquitin Cell Cycle Indicator
GO	Gene ontology
GOrilla	Gene Ontology enrichment analysis and visualization tool
GWAS	Genome-wide association studies
H2O	Water
HAVANA	Human and vertebrate analysis and annotation
HCl	Hydrochloric acid
HMGCR	HMG-CoA reductase
HRP	Horseradish peroxidase
IBD	Inflammatory bowel disease
IgG	Immunoglobulin G
IGMM	Institute of Genetics and Molecular Medicine
ISEMF	Intestinal subepithelial myofibroblast
IV	Intravenous
Kb	Kilo base pairs
kDa	Kilodaltons
LCL	Lymphoblastoid cell lines
LD	Linkage disequilibrium
lncRNA	Long non-coding RNA
LOH	Loss of heterozygosity

MAF	Minor allele frequency
MAP	MUTYH-associated polyposis
Mb	Mega base pairs
PMSF	Phenylmethanesulfonyl fluoride
MFC	Maximum fold change
M-MLV RT	Moloney Murine Leukemia Virus Reverse Transcriptase
MMR	DNA mismatch repair
MRC	Medical Research Council
mRNA	Messenger RNA
MSI	Microsatellite instability
Na3VO4	Sodium orthovanadate
NaF	Sodium fluoride
NaOAC	Sodium acetate
NaOH	Sodium hydroxide
NCBI	National Center for Biotechnology Information
NHS	National Health Service
NK-κB	Nuclear factor-κB
PBMC	Peripheral blood mononuclear cells
PBS	Phosphate buffered saline
PCR	Polymerase chain reaction
qRT-PCR	Quantitative real time PCR
R	R programming software
RIN	RNA integrity number
RNA	Ribonucleic acid
RNase	Ribonuclease
RNA-seq	RNA-sequencing
rRNA	Ribosomal RNA
RT-PCR	Reverse-transcription PCR
SAHSC	Scottish Academic Health Sciences Collaboration
SC	Scrambled siRNA
SCFA	Short-chain fatty acids
SDS	Sodium dodecyl sulfate

SEM	Standard error of mean
siRNA	Small interfering RNA
SMC	smooth muscle cell
SNP	Single nucleotide polymorphism
SOCCS	Scottish Colorectal Cancer Susceptibility
T/V	Trypsin/versene
TA	Transit-amplifying
TAE	Tris-Acetate EDTA
TCGA	The Cancer Genome Atlas
TEMED	N,N,N’,N’,tetramethyl-1-2-diaminomethane
TFBS	Transcription factor binding site
TSS	Transcription start site
UCSC	University of California, Santa Cruz
v/v	Volume per volume
VDR	Vitamin D receptor
VDRE	Vitamin D response elements
w/v	Weight per volume
WHO	World Health Organisation

TABLE OF CONTENTS

CHAPTER 2 Materials and Methods

CHAPTER 6 Cis-eQTL analysis of low-penetrance common genetic variants associated with colorectal cancer predisposition

CHAPTER 7 Identification of a novel indel polymorphism as the causal variant of the Xp22.2 colorectal cancer risk locus

CHAPTER 9 Functional characterisation of the gene-environment (circulating 25-hydroxyvitamin D) interaction at the 16q22.1 risk locus

CHAPTER 10 Summary and discussion

LIST OF TABLES AND FIGURES

Chapter 1

Introduction

1.1 Introduction

The understanding of the genetic predisposition to colorectal cancer (CRC) has progressed in the last decade with the advent of genome-wide association studies (GWAS). At least twenty-five common genetic variants have been established to be associated with CRC risk, invoking the common disease-common variant paradigm (Reich *et al*, 2001). However, the functional mechanisms by which they influence risk are not well-understood. Therefore, the investigation into these mechanisms has considerable relevance to understanding the aetiopathogenesis of this complex disease, which may ultimately lead to the discovery of novel therapeutic and preventative targets. The research presented in this thesis has systematically investigated whether these risk loci are associated with the baseline expression of nearby genes in tissue types relevant to colorectal cancer. Significant associations are prioritised and followed-up with functional assays to elucidate the causal molecular mechanisms.

The importance of delineating the molecular mechanisms that underlie CRC is underscored by the fact that colorectal cancer is a major health problem globally. In this introductory chapter, the incidence and burden of CRC is firstly discussed. The molecular events giving rise to CRC and its cell of origin are considered as these are pertinent issues that will influence the study design and the interpretation of various aspects of the gene expression analysis. A review of the risk factors associated with CRC is presented, including dietary/lifestyle factors, inflammation, the microbiome and genetic predisposition. This ranges from rare familial cancer syndromes to low-penetrance common susceptibility alleles, which form the impetus for this research. Though limited, the current understanding of the causal variants and mechanisms is described. As all of these risk loci reside within non-coding regions, it is thought that they confer risk by subtly influencing the regulation of gene expression and may also act as expression quantitative trait loci (eQTL). The use of eQTL analysis in complex disease traits and the functional annotation of CRC risk loci is therefore discussed. Finally, the aims of the project are presented and the experimental approaches are described.

1.2 Colorectal cancer: epidemiology and pathogenesis

1.2.1 Incidence and burden

Large bowel cancer, or colorectal cancer (CRC) is the third most common cause of cancer worldwide and the fourth biggest cause of cancer mortality, with nearly 1.4 million new cases diagnosed in 2012 (World Cancer Research Fund International; URL1.1). It is more common in the developed world, where the incidence is over two and a half times higher in developed countries compared to less developed ones. In the United Kingdom, there were on average 22,517 newly diagnosed cases of CRC per year in men, and 17,846 new cases in women during 2008-2010 (Office for National Statistics; URL1.2). It is estimated that 1 in 14 men and 1 in 19 women will develop CRC at some point in their lives (Cancer Research UK; URL1.3). In the United States, there has been a steady decline in the incidence of CRC in patients age 50 years or older during 1975-2010, but the opposite has been observed for young adults aged 20 to 49 years (Bailey *et al*, 2015).

Age-standardised rates suggest that bowel cancer is more common in industrialised countries with westernised societies. Global data from 2008 indicates that the WHO European region had the highest incidence of colorectal cancer followed by the WHO Americas region, whereas WHO African region had the lowest incidence. According to the World Bank income groups for countries, high income countries had considerably higher CRC incidence rates than any other income group, with nearly five times higher than the rate in low income countries (World Health Organisation; URL1.4). However, it should be noted that this manifestation of colorectal cancer burden may partly reflect longer life expectancy in developed populations, as well as better diagnostic and recording tools.

With earlier detection and improvements in treatment strategies, CRC mortality rates have decreased overall in the UK since the early 1970s. However, although it is a treatable disease with bowel surgery and adjuvant chemo-radiotherapy, the prognosis of CRC is still relatively poor. In 2012, there were 16,187 deaths from bowel cancer in the UK, of which 54% were men and 46% were women. (Cancer Research UK; URL1.5). Several factors are associated with higher risk of death, such as age, socio-economic deprivation, and most importantly, the stage of the cancer at diagnosis. In men, the five-year survival rate of 95% in stage I CRC falls dramatically to 7% in stage IV CRC. In women, five-year survival ranges from 100% at stage I to 8% at stage IV. There is compelling evidence that early detection and prevention by removal of premalignant polyps can reduce mortality, as indicated by randomised trials of population screening (Towler *et al*, 2007) and intensive surveillance of genetically defined high-risk groups (Jarvinen *et al*, 2000). An understanding of the disease aetiology and risk factors will not only allow risk modifications and preventative therapies, but also have an impact on targeted screening and treatment strategies.

1.2.2 Molecular genetics of colorectal cancer

Historically, CRC classification has been based only on clinical and pathological features. There is growing evidence that over the past decade that CRC is a heterogenous complex of diseases, where the molecular and genetic features of the tumour can determine prognosis and the response to therapeutic agents, in particular targeted therapy.

The sequence from the pre-malignant adenoma to carcinoma is well understood on a clinical level, and Vogelstein first described in his multistep genetic model that the accumulation of multiple mutations leads to the selective growth advantage that underlies tumourigenesis (Fearon and Vogelstein, 1990). In this model, the early loss or mutation of APC serve as the initiating event in adenoma formation, with at least seven distinct mutations required for carcinogenesis. Since then, genome-wide sequencing of CRC have calculated about 80 mutated genes per tumour, with less than 15 mutations considered to be true drivers (Wood *et al*, 2007). More recently, the alternative route of colon cancer carcinogenesis via serrated polyps have been described to account for 30% of CRC, where activating mutations of the mitogen-activated protein kinase pathway components *BRAF* or *KRAS* play a prominent role in this pathway (as reviewed in Bettington *et al*, 2013). Although the precise molecular events that lead to the development of CRC and its phenotypic changes are still not fully understood, there is now clear evidence for the presence of different subtypes of CRC.

There are at least three distinct molecular pathways that have been recognised to give rise to CRC. The chromosomal instability (CIN) pathway is defined by the accumulation of numerical (aneuploidy) or structural chromosomal abnormalities that result in karyotypic variability. It is the most common manifestation of genomic instability in CRC, occurring in approximately 70% of colorectal tumours (Lengauer *et al*, 1997), and is characterised by chromosomal rearrangements and loss-of-heterozygosity (LOH) at tumour suppressor gene loci. CIN tumours can also be discerned by the accumulation of mutations in specific oncogenes such as *APC, KRAS, PIK3CA, BRAF*, etc and tumour suppressor genes, but whether CIN creates the appropriate environment for the accumulation of these mutations or vice versa remains unclear (Pino *et al*, 2010).

The microsatellite instability (MSI) pathway is the other important pathway leading to genomic instability in CRC. It is characterised by genetic hypermutability caused by the dysfunction of DNA mismatch repair (MMR) genes. Deficiency in DNA repair gives rise to the accumulation of abnormalities in microsatellites, which are nucleotide repeat sequences of 1-6 base pairs that are prone to mutations due to the inability of DNA polymerases to bind these sequence motifs efficiently. As a result of insertions or deletions in coding regions, frameshift mutations occur with subsequent deleterious protein truncations. The DNA MMR system is inactivated either by germline mutations in MMR genes (as seen in the familial syndrome Lynch Syndrome), or epigenetically by gene promoter hypermethylation and silencing of *MLH1* in sporadic CRC (Herman *et al*, 1998; Veigel *et al*, 1998). More recently, the Cancer Genome Atlas Project (TCGA) demonstrated by whole-genome sequencing that a quarter of hypermutated tumours had somatic mismatch-repair gene and polymerase ε (POLE) mutations (Muzny *et al*, 2012).

Microsatellite instability in sporadic CRC that is related to hypermethylation and *MLH1* silencing is dependent on the third molecular pathway which is characterised by epigenetic instability as evident by the presence of widespread CpG island methylation (Toyota *et al*, 1999). The CpG Island Methylation Phenotype (CIMP) is associated with distinct genetic profiles, where CIMP1 is characterised by higher rates of MSI and *BRAF* mutations (Weisenberger *et al*, 2006; Shen *et al*, 2007), CIMP2 is associated with *KRAS* mutations, and CIMP-negative cases are enriched with *TP53* mutations (Shen *et al*, 2007; Hinoue *et al*, 2012).

CRC subtyping has also been addressed using gene-expression profiling in large patient cohorts, where molecular expression subtypes have not only been associated with different molecular pathways and cellular phenotypes, but also with prognosis and treatment responses (Salazar *et al*, 2011; De Sousa *et al*, 2013; Sadanandam *et al*, 2013).

1.2.3 Cell of origin of colorectal cancer

The epithelial layer of the human large intestine consists of a single sheet of columnar epithelial cells, which form crypt-like invaginations into the lamina propria connective tissue to form the functional units of the colon. The four major terminally differentiated epithelial cell types in the colonic crypt are known as the enterocytes (absorptive cells), the goblet cells (mucus-secreting), the enteroendocrine cells (peptide hormone-secreting), and the recently characterised tuft cells (opioid and prostaglandin-secreting) (Gerbe *et al*, 2013) (Figure 1.1). The organisation, architecture, differentiation and homeostasis of the crypt component cells are pivotal to the normal functioning of the colonic epithelium, and are thought to be maintained by the gene expression gradients of key signalling molecules along the vertical crypt axis, mediated by autocrine and paracrine pathways that arise from epithelial-mesenchymal interactions (Figure 1.2). The key signalling pathways implicated are those of Wnt (Korinek *et al*, 1998; Pinto *et al*, 2003; Sansom *et al*, 2004), EphB/Ephrin B (Batlle *et al*, 2002), Notch (Jensen *et al*, 2000; van Es *et al*, 2005), BMP (He *et al*, 2004; Kosinski *et al*, 2007) and Hedgehog (Madison *et al*, 2005).

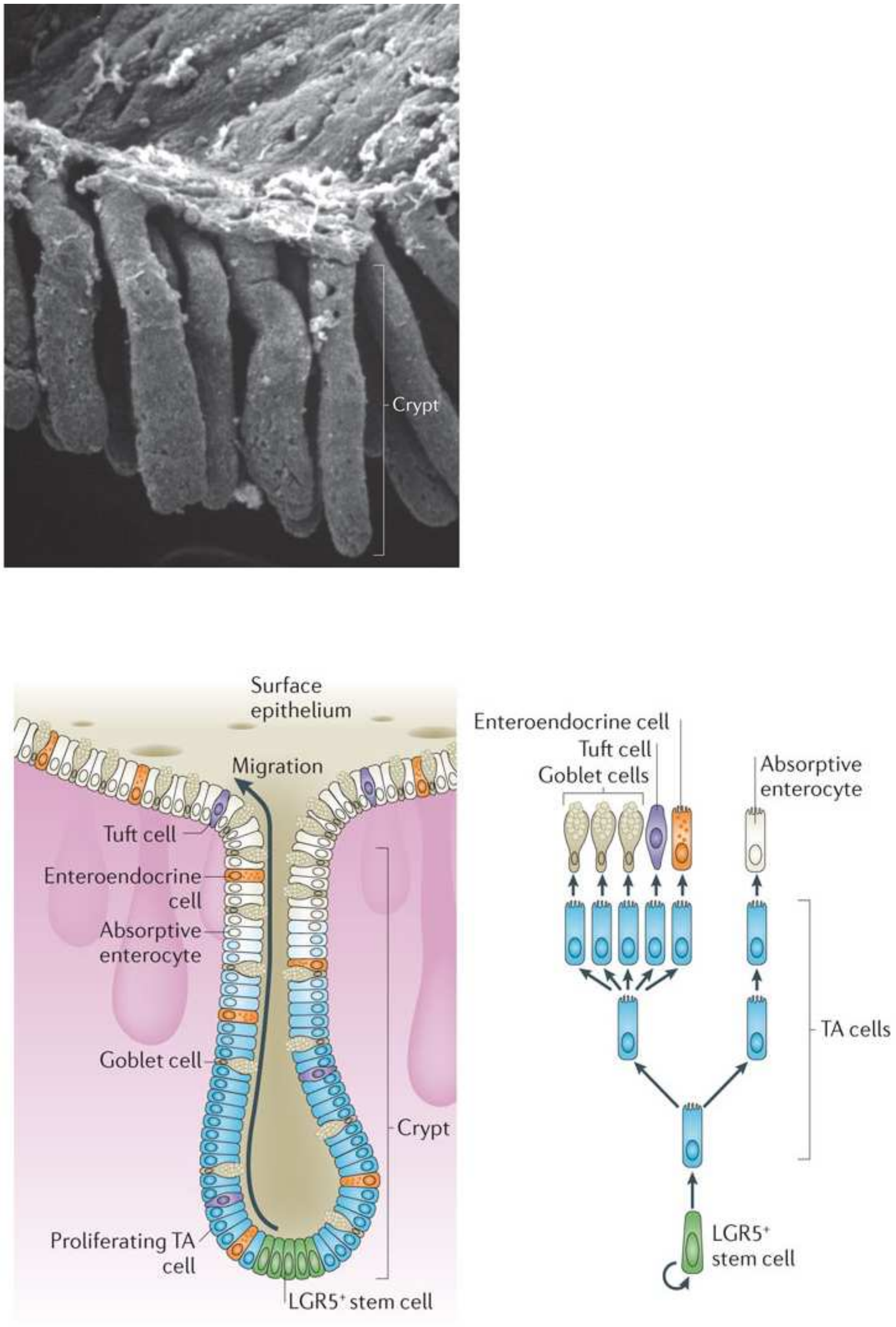


Figure 1.1 In the colon (scanning electron micrograph in top panel), LGR5$^+$ stem cells at the crypt base generate rapidly proliferating TA (transit-amplifying) cells in the lower half of the crypt (bottom left panel). TA cells subsequently differentiate into the mature lineages of the surface epithelium (enterocytes, goblet cells, enteroendocrine cells and tuft cells), as shown in the lineage tree (bottom right panel). Epithelial turnover occurs every 5–7 days. Adapted from Barker, 2014.

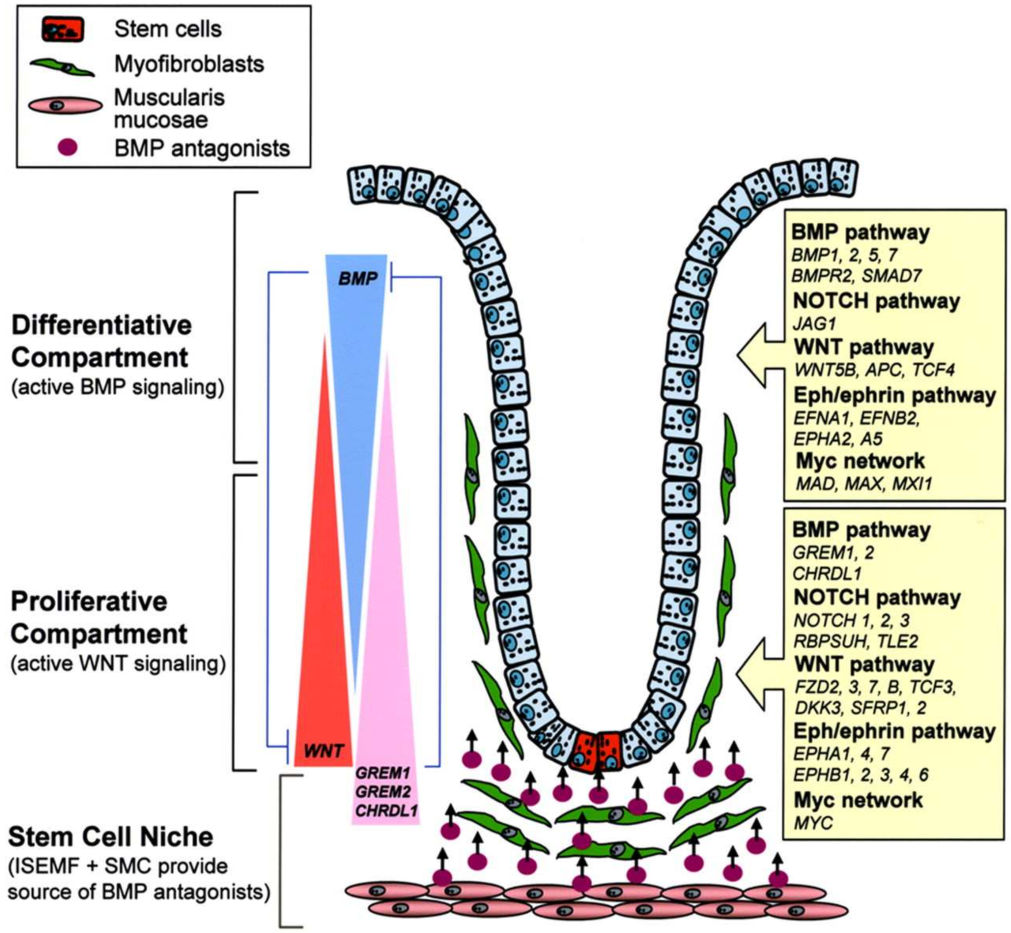


Figure 1.2 Signalling pathways that are involved in the regulation of homeostasis and determination of cell fate that are coupled to position along the vertical crypt axis of the epithelium. ISEMF, intestinal subepithelial myofibroblast; SMC, smooth muscle cell. (Kosinski *et al*, 2007)

There is a high rate of cell death and rapid turnover due to persistent abrasion from the luminal contents, which imposes a requirement for daily self-renewal driven by small populations of adult stem cells. The evidence points towards a stem-cell population that resides at the base of the crypt within the stem-cell niche, formed by the stem cells themselves and surrounding mesenchymal cells, the intestinal subepithelial myofibroblasts. Crucially, lineage-tracing experiments in mice using inducible stem-cell markers have confirmed monoclonal conversion and multipotentiality in the intestinal crypts, where the stem cell marker LGR5$^+$ crypt base columnar cells was shown to generate all epithelial lineages over a 60-day period (Barker *et al*, 2007). CD24$^+$ and KIT$^+$ goblet cells that are in close proximity to LGR5$^+$ stem cells at the crypt base have been identified as probable niche components (Rothernberg *et al*, 2012), but the major source of Wnt in the colon has yet to be identified.

Although genetic/epigenetic lesions are widely accepted to have a major role in determining tumour phenotype, it is also thought that cancers of distinct subtypes may derive from different 'cells of origin' leading to inter- and intra-tumoural heterogeneity (Visvader, 2011). In studies of colorectal cancer, there is accumulating evidence that supports a bottom-up theory of cancer origin, as the ability of stem cells to indefinitely self-renew while generating new functional epithelia makes them prime candidates for accumulating sequential genetic or epigenetic mutations that promote oncogenesis. Two distinct crypt stem cells have been identified as the cells of origin of intestinal cancers using an in-vivo targeting approach in mouse models that involves lineage tracing of cells as they undergo transformation. APC deletion in long-lived LGR5$^+$ stem cells but not in short-lived transit-amplifying cells revealed that intestinal cancer in mice originates from crypt stem cells (Barker *et al*, 2009). This target cell is also marked by PROM1 (Zhu *et al*, 2009). A BMI1$^+$ stem cell located in the +4 or +5 position from the base of the crypt and therefore distinct from the LGR5$^+$ stem cell was also shown to be susceptible to tumourigenesis by deregulated Wnt signalling.

In contrast, the top-down hypothesis of intestinal cancer postulates that any cell in the normal cellular hierarchy with proliferative capacity could also serve as a cell of origin of cancer, if it acquires mutations that re-instigate self-renewal capacity and prevent differentiation to a post-mitotic state. Supporting this paradigm are several recent transgenic mouse model studies that implicate distinct mechanisms involving non-stem cells. Schwitalla *et al* demonstrated that the combination of β-catenin activation and NF-κB signaling can convert LGR5^{-} cells into LGR5^{+} stem cells that give rise to intestinal neoplasms, exemplifying the concept of cell-type plasticity and bidirectional conversion that results in the dedifferentiation of non-stem cells, allowing them to act as tumour progenitors (Schwitalla *et al*, 2013). Consistent with this study, a lineage-tracing study of tuft cells demonstrated relative quiescence and longevity of a small number of DCLK1^{+} cells, which converted into potent cancer-initiating cells when subjected to a combination of *APC* loss and an inflammatory stimulus (Westphalen *et al*; 2014). Non-inflammatory processes have also been implicated; a recent mouse model of hereditary mixed polyposis syndrome (HMPS) showed that the aberrant epithelial expression of *GREM1* can promote the persistence and/or reacquisition of stem cell and tumour-initiating properties in *LGR5*$^{-}$ progenitor cells that have exited the stem cell niche by disrupting homeostatic intestinal morphogen gradients (Davis *et al*, 2015). In all likelihood, the cellular and molecular mechanisms underlying both hypotheses are not mutually exclusive and most probably act together as well as interact with extrinsic mechanisms such as the stromal micro-environment to determine tumour histopathology and behaviour.

1.3 Colorectal cancer: risk factors

Colorectal cancer typically develops over many years, with a multifactorial aetiology that involves environmental factors, genetic susceptibility and their interactions. It occurs more frequently in the distal large bowel (descending colon and rectum) compared to the more proximal regions of the large intestine (Rabaneck *et al*, 2003), which might reflect differences in the luminal environment and inherent cellular variation between these gut compartments. The risk factors that increases ones susceptibility to the disease have been extensively reviewed elsewhere (Raskov *et al*, 2014; Tenesa *et al*, 2009; Terzić *et al*, 2010; Louis *et al*, 2014) but the main themes will be presented here.

1.3.1 Dietary and lifestyle risk factors

The higher incidence of CRC in developed countries is suggestive of a contribution from environmental factors, broadly defined to include a wide range of cultural, lifestyle and dietary practices. This is evident from early studies of migrants from low to high incidence countries, who attain cancer incidence rates similar to those of their adopted country within a single generation (as reviewed by Boyle *et al*, 2000). Supporting this further are the rapidly increasing incidence rates in developed and westernised Asian countries with previously low rates, possibly reflecting lifestyle changes as well as gene-environment interactions (as reviewed in Sung *et al*, 2005).

Although there is little doubt that diet contributes to the development of CRC, studies that accurately examine the relationship between a specific food item and cancer are difficult to design, not least because the dietary assessment methods are inherently subjected to recall bias. Nevertheless, there are several dietary elements that have been shown to be linked to CRC.

A high intake of dietary fibre, in particular cereal fibre and whole grains, has been associated with a reduced risk of colorectal cancer (Aune *et al*, 2011). The partial or total fermentation of fibre in the colon leads to the production of short chain fatty acids such as butyrate, and it is thought that these play a pivotal role in maintaining normal colonic function and preventing disease by reducing the intraluminal pH, decreasing the conversion of bile acids to secondary bile acids (Birkett *et al*, 1996), and more importantly, exerting anti-proliferative properties and tumour-suppressive effects (Leonel *et al*, 2012; Fung *et al*, 2013). Dietary fibres also have the effect of diluting stool contents, bulking and increasing the frequency of bowel movements, reducing the concentration and contact time of carcinogens (Anderson *et al*, 2009).

Numerous prospective studies have linked meat consumption, in particular red meat and processed meat, to a higher risk of CRC (Larsson *et al*, 2006; Chan *et al*, 2011). A dose-response meta-analysis of epidemiological studies suggest that there was a 24% increase in CRC risk for an increase of 120g/day of red meat and a 36% increase in risk for 30g/day of processed meat (Norat *et al*, 2002). It is postulated that the haem iron in red meat has a catalytic effect on the endogenous formation of carcinogenic N-nitroso compounds and the formation of cytotoxic and genotoxic aldehydes (Bastide *et al*, 2011). The nitrites found in processed meat are also converted to N-nitroso compounds in the bowel.

Early epidemiological observations showed that the incidence and death rates of CRC were lower among individuals living in southern latitudes with relatively higher sunlight exposure, than among those living at northern latitudes (Garland *et al*, 1980). Because exposure to ultraviolet-B sunlight leads to the production of vitamin D, it was hypothesised that the variation in vitamin D levels might account for this association. This hypothesis have since been tested in various ways, including association studies with annual solar radiation levels (Mizoue *et al*, 2004), seasonality (Robsahm *et al*, 2004), dietary vitamin D intake (Grant *et al*, 2004; Giovannucci *et al*, 2005; Touvier *et al*, 2011), pre-diagnostic circulating vitamin D levels (Garland *et al*, 1989; Tangrea *et al*, 1997; Feskanich *et al*, 2004), genetic polymorphisms in the vitamin D receptor (Wong *et al*, 2003; Park *et al*, 2005, Touvier *et al*, 2011), and a composite score of multiple vitamin D predictors including skin pigmentation, region of residence, dietary intake, body mass index and physical activity (Giovannucci *et al*, 2005). Although establishing a causal relationship between CRC incidence and vitamin D is challenging because environmental risk factors associated with CRC may also be associated with vitamin D deficiency (e.g. co-causality with physical activity), all the epidemiological findings point towards hypovitaminosis D as a risk factor for developing CRC, with biological data to suggest that the vitamin D pathway activation induces cellular differentiation and inhibits proliferation, invasiveness, angiogenesis and metastatic potential (as reviewed by Peterlik *et al*, 2004).

There is also some evidence of an association between total energy intake and the risk of developing CRC. However, this is relationship is likely to be indirect and may be dependent on other factors, such as physical activity (as reviewed by Wiseman, 2008). Exercise appears to have a dose-response reduction in the rate of CRC, and it is postulated that the increase in insulin-like growth factor-binding protein and the reduction of prostaglandins may be the mechanism by which exercise provides this protective effect. Other non-dietary factors have also been associated with increasing the risk of CRC. Tobacco smoking doubles the risk of colorectal adenomas (Botteri *et al*, 2008), and other cohort studies have found that alcohol intake increases the risk of CRC (Moskal *et al*, 2007).

1.3.2 Inflammation

Inflammatory bowel disease is a major risk factor for developing colorectal cancer. Over 20% of patients with inflammatory bowel disease develop colitis-associated cancer (CAC) within 30 years of disease onset, a subtype of CRC that is associated with a high mortality of >50% (Lakatos *et al*, 2008). Although CAC is thought of as a distinct subtype of colorectal cancer, there are similarities between CAC and other types of CRC that develop without any signs of overt inflammatory disease. The essential stages of cancer development such as aberrant crypt foci, polyps, adenomas and carcinomas, as well as common genetic and signalling pathways such as those involving Wnt, β-catenin, KRAS, p53, TGF-β, and DNA mismatch repair, are similar between CAC and sporadic CRC (as reviewed by Terzić *et al*, 2010). Furthermore, sporadic CRC display inflammatory infiltration and increased pro-inflammatory cytokine expression (Clevers, 2004; Atreya *et al*, 2008). There is evidence from numerous observational studies that non-steroidal anti-inflammatory drugs such as sulindac, celecoxib and aspirin may have chemopreventative effects, and it is thought that these compounds mediate risk reduction by modulating cyclooxygenase (COX) enzymatic activity and the nuclear factor-κB (NK-κB) pathway (Yamamoto *et al*, 1999; Larsson *et al*, 2006, Flossmann *et al*, 2007; Chan *et al*, 2007; Arber *et al*, 2006, Meyskens *et al*, 2008; Half *et al*, 2009). Of these agents, the evidence for aspirin is most convincing, with a large randomised controlled trial showing a risk reduction in high-risk individuals taking low-dose aspirin (Burn *et al*, 2011).

1.3.3 The microbiome

There is emerging interest in the role of the microbiota in the initiation and progression of CRC. Microbiome changes that have been reported to be observed in CRC patients include *S. bovis* (as reviewed in Burnett-Hartman *et al*, 2008), Streptococcus spp. (Wang *et al*, 2012), *Escherichia coli* (Bonnet *et al*, 2014), *Fusobacterium nucleatum* (Castellarin *et al*, 2012; Kostic *et al*, 2012), Clostridium (Scanlan *et al*, 2008), Bacteroides (Wang *et al*, 2012) and various butyrate-producing bacteria (Balamurugan *et al*, 2008; Wang *et al*, 2012). Apart from these observed associations, experimental animal studies support a direct influence of the gut microbiota on tumour formation (Dove *et al*, 1997; Arthur *et al*, 2012) that is inter-dependent with the host inflammatory response (Arthur *et al*, 2014; Boleji *et al*, 2010).

Bacterial metabolism in the colon is fermentative and also utilises anaerobic respiration. As alluded to in the previous section, undigested dietary components and endogenous products are fermented by the anaerobic microbial community to produce an extraordinarily wide range of metabolites. The major fermentation products are organic acids, in particular the short-chain fatty acids (SCFA) acetate, propionate and butyrates. Aside from providing an energy source to gut epithelial cells, SCFA have been shown to regulate colonic regulatory T-cells (Smith *et al*, 2013), downregulate pro-inflammatory cytokines in colonic macrophages by inhibiting the activity of histone deacetylases (Chang *et al*, 2014), selectively induce apoptosis of CRC cells (Buda *et al*, 2003; Clarke *et al*, 2008), and maintain intestinal homeostasis. Prominent butyrate-producing species indicates healthy, diverse microbiota, and maintains favourable conditions for a stable and healthy gut community. By contrast, dysbiosis is characterized by a reduction in microbial diversity and an increase in pro-inflammatory, pathogenic species, which can be caused by a poor diet, antimicrobial therapy or genetic predisposition.

The microbial communities that inhabit our gastrointestinal tract are tractable environmental factors that we are exposed to continuously, and it has become increasingly clear that the collective activities of the resident gut microbiota and their metabolic products plays a role in the development of CRC (as reviewed by Schwabe *et al*, 2013; Louis *et al*, 2014). Hence, it is likely that there is a multifaceted relationship between diet and microbial metabolism that promotes CRC via pro-inflammatory interactions with host intestinal cells.

1.3.4 Genetic heritability

It has long been known that inherited susceptibility plays an important role in the predisposition to CRC. The earliest evidence for this came from epidemiological studies in the fifties that showed a 2-3 fold increase in risk in first degree relatives of CRC patients (Johns *et al*, 2001). Analysis of phenotype concordance in twins estimates the heritability of colorectal cancer on the liability scale to be around 0.35 (Lichtenstein *et al*, 2000). Until recently, our understanding of the hereditary component was based on rare, high-penetrance mutations in a few genes, such as *APC*, mismatch repair (MMR) genes, *SMAD4*, and *MUTYH*. Despite the large effects of these rare variants, their low allele frequency means their overall contribution to disease burden is small (Foulkes *et al*, 2008). Statistical modelling of the pattern of familial occurrence of colorectal cancer after exclusion of known high-risk genes suggest that the remaining genetic heritability is likely to be polygenic with the co-inheritance of multiple genetic variants, each with a modest individual effect, causing a wide range of risk in the population (Figure 1.3). Rare, moderately-penetrant risk alleles (MAF<2%; relative risks>2.0) and common, low-penetrance alleles (MAF>10%; OR<1.5) are likely to occur as a continuum, and extensive efforts are underway to comprehensively identify these susceptibility variants.

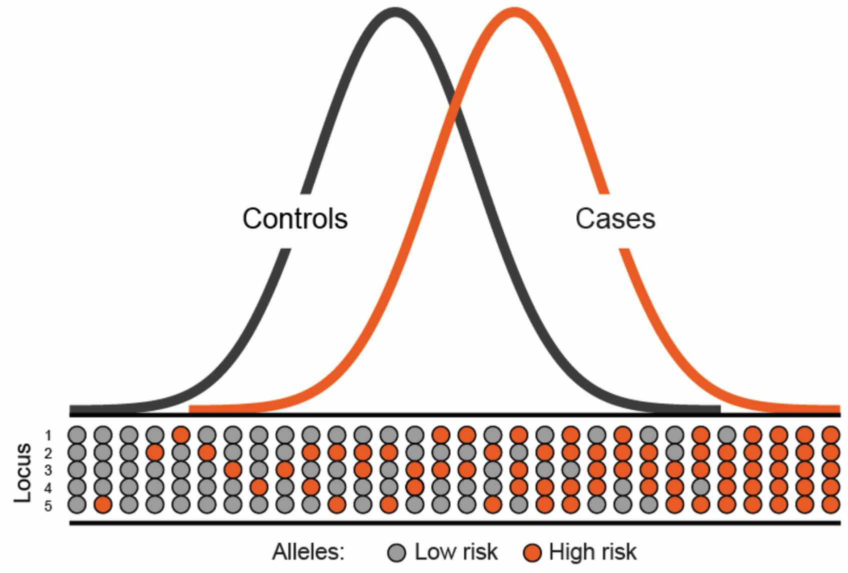


Figure 1.3. Polygenic model of disease susceptibility. Cases have a shift towards a higher number of high risk alleles. Adapted from Whiffin *et al*, 2014.

1.3.4.1 Very rare, high-penetrance familial colorectal cancer syndromes

Hereditary CRC, where a clear genetic basis for the disease has been defined, accounts for 4-6% of colon cancer incidence (Rustgi, 2007). Family-based genetic linkage and positional cloning studies have led to the identification of numerous CRC susceptibility genes (Table 1.1). The two major Mendelian cancer syndromes that account for the vast majority of hereditary CRC cases include Familial Adenomatous Polyposis (FAP) and Lynch Syndrome.

GENE(S)	SYNDROME	RISK IN MUTATION CARRIERS	MODE OF INHERITANCE
APC	FAP	90% by age 45	Dominant
Mismatch repair genes	Lynch Syndrome	40%–80% by age 75	Dominant
MUTYH	MYH-associated polyposis	35%–53%	Recessive
SMAD4/BMPR1A	Juvenile Polyposis syndrome	17%–68% by age 60	Dominant
STK11	Peutz-Jeghers syndrome	39% by age 70	Dominant
PTEN	Cowden syndrome	15% lifetime risk	Dominant
POLD1/POLE	Oligopolyposis		Dominant

Table 1.1 Familial CRC syndromes and the associated high-penetrance gene mutations. Adapted from Whiffin *et al*, 2014.

Familial adenomatous polyposis (FAP) is characterized by the development of hundreds to thousands of benign adenomatous polyps that carpet the colon and rectum of affected individuals. These polyps usually appear during the second or third decade of life. If the colon is not removed, cancer will inevitably develop in all FAP patients, with an average age of colon cancer development of 39 years (Wills *et al*, 2002). It is inherited as an autosomal dominant disease, with a population incidence of approximately 1 in 8000 (Bisgaard *et al*, 1994). Germline mutations of the *APC* gene on chromosome 5q21 is responsible for over 95% of affected families. *APC* mutations achieve near 100% penetrance, although there is marked variation in phenotypic expression of the disease. Extracolonic tumours also occur and include small bowel, gastric and periampullary tumours, adrenal adenomas and carcinomas, and thyroid carcinomas. Other associated lesions include desmoid tumours and congenital hypertrophy of the retinal pigment epithelium (CHRPE) (Lynch *et al*, 1998).

The gene product of the intact *APC* gene functions as a tumour suppressor. It is a negative regulator in the Wnt signalling pathway (Goss *et al*, 2000), where it binds to and phosphorylates soluble beta-catenin leading to its cytosolic degradation. In FAP, the loss of functioning APC protein leads to the unregulated translocation and accumulation of beta-catenin in the cell nucleus, where it interacts with TCF/LEF transcription factors to constitutively activate the transcription of many gene targets including *MYC*, *CCND1*, *CD44* and *BMP4* (Tetsu *et al*, 1999; He *et al*, 1998; van de Wetering *et al*, 2002). The loss of wild-type *APC* may also affect tumourigenesis via other mechanisms such as the regulation of cell migration (Kawasaki *et al*, 2003; Sansom *et al*, 2004) and the organisation of the actin cytoskeletal network (Watanabe *et al*, 2004).

Lynch Syndrome is an autosomal dominant disorder and is the most common familial CRC syndrome, accounting for 2% - 3% of all CRC cases (Lynch *et al*, 2000). Without a distinct polyposis phenotype, a detailed family history becomes critical in the detection of Lynch Syndrome families. Lynch Syndrome tumours tend to display an earlier onset than sporadic colon cancers and are more likely to occur in the proximal colon (Lynch *et al*, 2009). Apart from CRC, Lynch Syndrome families also see a predisposition for extra-colonic cancers, most notably endometrial cancers, as well as cancers of the ovary, small intestine, stomach, hepatobiliary tract, urinary tract and brain. Germline mutations in one of the MMR genes are responsible for this susceptibility disorder, and confers a lifetime risk of CRC and endometrial cancer of 60-80% and 40-60% respectively (Meyer *et al*, 2009). The penetrance has been observed to be significantly greater in males than females (74% vs 30%) but the risk of endometrial cancer exceeded that for CRC in females (42%) (Dunlop *et al*, 1997), suggesting that there are gender-specific modifiers of risk.

There has been a total of seven genes identified as members of the MMR family, with the majority of Lynch Syndrome families having mutations in either *MSH2* or *MLH1* (Liu *et al*, 1996; Mitchell *et al*, 2002). Mutations of *MSH6* have been identified in a small minority of cases (Kolodner *et al*, 1999), while rare mutations in *PMS1* and *PMS2* have been reported (Nicolaides *et al*, 1994; Worthley *et al*, 2005). Lynch Syndrome is usually caused by the inheritance of one mutant MMR allele and loss of heterozygosity at the remaining wild-type allele. This leads to a mutator phenotype where cells accumulate further mutations at an amplified rate, increasing the probability of mutations in other proto-oncogenes and tumour suppressors. The mutator phenotype manifests as a specific genomic instability event at small repeated sequences in DNA called microsatellite instability (Liu *et al*, 1996), as the MMR system is less effective in correcting the slippage error of DNA polymerase at highly repetitive regions of DNA. There are also other non-repair functions of the MMR pathway that may contribute to tumourigenesis, such as the activation of cell cycle checkpoints and apoptosis in response to DNA-damaging agents and the maintenance of homologous recombination fidelity (as reviewed by Heinen, 2011). Evidence from studies of mouse models suggest that although the development of a mutator phenotype is sufficient to drive tumourigenesis, the ability of MMR-defective cells to survive under conditions of increased damage may accelerate the process (Lin *et al*, 2004; Yang *et al*; 2004).

In the last decade, families with an attenuated FAP-like phenotype that do not appear to carry any germline mutation in *APC* have been described. Over 25% of these patients carried germline biallelic mutations in the base-excision repair gene *MUTYH* (Al-Tassan *et al*, 2002; Jones *et al*, 2002; Sieber *et al*, 2003; Farrington *et al*; 2005), and this form of recessive hereditary cancer has been termed *MUTYH*-associated polyposis (MAP). More recently, specific germline exonuclease domain mutations in polymerase proofreading genes *POLD1* and *POLE* have also been identified to be causative for the development of multiple colorectal adenomas and CRC (Palles *et al*, 2013). Collectively, the MMR defects of Lynch Syndrome, base excision repair defects that cause MAP, and the mutations in proofreading genes emphasise the critical role of replication errors and coupled repair of base pair-level mutations in the predisposition to CRC.

Rarer mutations in other genes associated with hereditary CRC include those in *STK11* (Peutz-Jeghers syndrome), *PTEN* (Cowden's disease) and *BMPR1A/SMAD4* (Juvenile Polyposis) (Ngeow *et al*, 2013), where CRC risk is mediated through the development of hamartomas or mixed polyps. In comparison to the gatekeeper function of the *APC* gene and the caretaker roles of the mismatch repair and *MUTYH* genes, these genes are believed to create an epithelial milieu at risk for neoplastic development and have been dubbed 'landscaper' genes (Kinzler *et al*, 1998), highlighting the various signalling pathways that contributes to the formation of cancer.

1.3.4.2 Rare, moderately-penetrant risk variants

Candidate gene resequencing studies in affected families have been the mainstay of the methodologies used to identify this subgroup of risk variants. As these approaches relied on *a priori* knowledge, their success has been hampered by our limited knowledge of tumour biology. Rare successes of this approach include the discovery of the missense variant (APC I1307K) that is present in ~6% of Ashkenazi Jews (Laken *et al*, 1997). The APC I1307K T>A variant creates a small hypermutable region that appears to increase replication errors in *APC*, increasing the risk of CRC by approximately two-fold.

With the advent of large-scale exome sequencing studies in recent years, exome arrays have been specifically designed to allow exome-wide systematic interrogation of coding variants with putative detrimental functional consequences. In a large unrelated case-control study, four novel coding variants were identified to be associated with CRC risk (Timofeeva *et al*, submitted). However, the minor alleles of these variants are common (MAF 0.09-0.50) with modest effect sizes (OR 1.08-1.15). No rare alleles (MAF<0.05) of moderate effect were identified, despite adequate power to detect such effect sizes. This is contrary to the expectation that coding sequence variants with putative deleterious effects might have a more profound impact on risk. This suggests that rare genetic variation of moderate penetrance are likely to segregate in families, and that exome and genome sequencing of trios and families may be a better strategy to identify these variants.

1.3.4.3 Common genetic risk variants

A substantial proportion of the remaining heritable risk is likely to be accounted for by numerous low-penetrance genetic variants, each with a relatively high frequency in the population, as described in the “common disease-common variant” hypothesis. This model posits that if a heritable disease is common in the population, then the genetic contributors will also be common in the population. However, even though the contribution of an individual variant to the overall inherited susceptibility of a disease may be relatively large, the penetrance of these variants will be very small, which would explain why these variants rarely cause multiple cases in families and hence are not detectable through genetic linkage studies.

Until mid-2007, no common variants contributing to the heritability of colorectal cancer risk had been successfully identified and consistently replicated. In the last decade, genome-wide association studies (GWAS) have provided a new conceptual framework in the search for the genetic basis of CRC. By exploiting the non-random coinheritance of genetic variants (linkage disequilibrium [LD]), these studies utilise single nucleotide polymorphism (SNP) “tags” for haplotypes to representatively assay the entire genome. As the genome is screened without any prior hypothesis for specific regions, genes, or variants thereof, GWAS are regarded as “agnostic” or hypothesis-generating, rather than hypothesis-driven. The last decade has seen the assembly of large well-characterised case-control series with sufficient power to account for the large number of statistical tests performed and detect small effect sizes. Facilitated by technological advances and cost-reduction in high-density reproducible genotyping platforms, over twenty common low-penetrance variants have since been identified to be associated with CRC, all of which have been validated in multiple case-control cohorts from various populations (Table 1.2).

tagSNP	Locus	SNP position	MAF	Effect size: OR (95% CI)	Reference
rs10911251	1q25.3	chr1:183081194	0.39	1.09 (1.06-1.13)	Peters *et al*, 2013
rs6687758	1q41	chr1:222164948	0.22	1.09 (1.06–1.12)	Houlston *et al*, 2010
rs6691170	1q41	chr1:222045446	0.26	1.06 (1.03-1.09)	Houlston *et al*, 2010
rs10936599	3q26.2	chr3:169492101	0.30	0.93 (0.91–0.96)	Houlston *et al*, 2010
rs1321311	6p21.2	chr6:36622900	0.25	1.10 (1.07-1.13)	Dunlop *et al*, 2012
rs16892766	8q23.3	chr8:117630683	0.07	1.27 (1.20-1.34)*	Tomlinson *et al*, 2008
rs6983267	8q24.21	chr8:128413305	0.49	1.17 (1.12-1.23)*	Zanke *et al*, 2007; Tomlinson *et al*, 2007
rs1035209	10q24.2	chr10:101345366	0.14	1.12 (1.08-1.16)	Whiffin *et al*, 2014
rs10795668	10p14	chr10:8701219	0.33	0.87 (0.83-0.91)*	Tomlinson *et al*, 2008
rs3802842	11q23.1	chr11:111171709	0.29	1.11 (1.08-1.15)	Tenesa *et al*, 2008
rs3824999	11q13.4	chr11:74345550	0.38	1.08 (1.05-1.10)	Dunlop *et al*, 2012
rs3217810	12p13.32	chr12:4388271	0.06	1.20 (1.12-1.28)	Peters *et al*, 2013
rs11169552	12q13.13	chr12:51155663	0.24	0.92 (0.90–0.95)	Houlston *et al*, 2010
rs7136702	12q13.13	chr12:50880216	0.46	1.06 (1.04–1.08)	Houlston *et al*, 2010
rs1957636	14q22.2	chr14:54560018	0.43	1.08 (1.06-1.11)	Tomlinson *et al*, 2011
rs4444235	14q22.2	chr14:54410919	0.46	1.11 (1.08-1.15)	Houlston *et al*, 2008
rs11632715	15q13.3	chr15:33004247	0.47	1.12 (1.08-1.16)	Tomlinson *et al*, 2011
rs16969681	15q13.3	chr15:32993111	0.18	1.18 (1.11-1.25)	Tomlinson *et al*, 2011
rs9929218	16q22.1	chr16:68820946	0.29	0.90 (0.87-0.94)	Houlston *et al*, 2008
rs4939827	18q21.1	chr18:46453463	0.48	0.86 (0.79–0.92)	Broderick *et al*, 2007
rs10411210	19q13.11	chr19:33532300	0.10	0.83 (0.78-0.88)	Houlston *et al*, 2008
rs4813802	20p12.3	chr20:6699595	0.25	1.09 (1.06-1.12)	Tomlinson *et al*, 2011
rs4925386	20q13.33	chr20:60921044	0.41	0.93 (0.91–0.95)	Houlston *et al*, 2010
rs961253	20p12.3	chr20:6404281	0.35	1.12 (1.08-1.16)	Houlston *et al*, 2008
rs5934683	Xp22.2	chrX:9751474	0.37	1.07 (1.04-1.10)	Dunlop *et al*, 2012

Table 1.2 Twenty-five single-nucleotide polymorphisms that are associated with CRC as identified from GWAS. * denotes OR for heterozygotes are presented when the OR per allele is not calculated.

The first GWAS for CRC were carried out in Scotland (Zanke *et al*, 2007; Tenesa *et al*, 2008), England (Tomlinson *et al*, 2007; Broderick *et al*, 2008), and Canada (Zanke *et al*, 2007). These studies utilised a primary phase that involved modest sample sizes (~1000 cases and 1000 controls genotyped for ~0.5 million tagging SNPs), followed by larger validation phases of those SNPs with the strongest signals of association. Although the six initial genetic variant associations with CRC were highly significant and passed the stringent threshold of multiple-testing, the effect size of these variants were modest at best (odds ratio ≈ 1.2). Consequently, the power to detect the effects of such loci was modest, with the likelihood of discovery being highly sensitive to small chance differences in genotype frequencies. Hence, it is thought that many more CRC loci of similar or smaller effect size may exist, prompting further large-scale collaborative efforts to discover new risk variants that may not be easily discoverable owing to small effect sizes and/or low risk allele frequencies. Meta-analyses of all initial UK GWAS data (Houlston *et al*, 2008) and further case-control sets (Houlston *et al*, 2010, Dunlop *et al*, 2012; Whiffin *et al*, 2014) revealed fourteen further risk loci with even smaller effect sizes (odds ratio ≤ 1.1) than those that had been detected previously. Of note, an X-linked locus at Xp22.2 was associated with CRC, and represents the first evidence for the role of X-chromosome variation in the predisposition to a non-sex specific cancer (Dunlop *et al*, 2012).

Other new variants have also been discovered by modifying the traditional GWAS approaches. A candidate-gene based fine-mapping study was able to identify new predisposition tagging SNPs, as well as deconvolute the tagSNP association at the *GREM1* locus, demonstrating that the original rs4779584 SNP was a synthetic association tagging two independent functional SNPs (Tomlinson *et al*, 2011). To increase sample size and statistical power, a meta-analysis included colorectal adenoma cases based on the knowledge that adenomas are well-defined CRC precursors and hence share a similar aetiology with the malignant phenotype (Peters *et al*, 2013). However, two out of their four putative associations failed to be replicated in another meta-analysis that shared an overlapping sample set (Whiffin *et al*, 2014). As these studies relied on different imputation references (HapMap 30 trios in Peters *et al*, 2013 versus 1000 Genomes Project in Whiffin *et al*, 2014) to recover the genotypes of this two SNPs, the failure of replication is likely to reflect discrepancies in imputation, a key issue pertinent to the later GWAS studies. Although imputation with publicly available, well-catalogued deep-sequencing data is a highly useful and cost-effective measure to complement genotyping arrays, technical validation of imputation fidelity by sequencing is paramount to avoid spurious results.

GWAS have proved to be a powerful approach in identifying common, low penetrance susceptibility loci for CRC without prior knowledge of disease pathways. Although each individual risk allele confers only a small relative risk, the SNPs are common and hence contribute significantly to the overall incidence of CRC. Furthermore, the accrual of risk variants in an individual also impacts significantly on an individual's risk of developing CRC (Figure 1.4), and may allow the identification of higher-risk individuals in the general population who might benefit from earlier screening (Dunlop *et al*, 2012; Lubbe *et al*, 2012). Although the collective risk conferred by currently identified common variation explains only ~2% of colorectal cancer, this estimate is likely to be conservative for several reasons. Firstly, the effect of the causal variant(s) at each locus is expected to be larger than the association detected by the tagging variant. As evidenced by the 14q22 association, multiple risk variants may exist at each locus, including low-frequency variants with significantly larger effects (Tomlinson *et al*, 2011). Secondly, the interactions of these variants with epigenetic regulation or environmental factors may lead to a greater increase in disease risk. Epistatic interactions between these low-penetrance variants could in theory result in a larger impact on CRC risk (Zuk *et al*, 2012), however, the evidence to date suggests that the effects of most risk loci appear to be independent. Aside from effect underestimation of established risk loci, the remaining heritable susceptibility may also be embodied in a multitude of common risk alleles with even smaller effect sizes that are yet to be identified. This is evidenced by larger and more highly-powered GWAS in breast (Michailidou *et al*, 2013) and prostate (Eeles *et al*, 2013) cancer, identifying 41 and 23 novel risk loci respectively. Furthermore, structural variation such as indels and copy number variation that are likely to play a role in disease predisposition are not optimally captured by commerical GWAS arrays, and may account for some of the missing heritability that have eluded GWAS efforts thus far. A new generation of studies involving exome and whole-genome sequencing, as well as gene-environment interactions are hence underway to improve our understanding in the inherited predisposition to CRC.

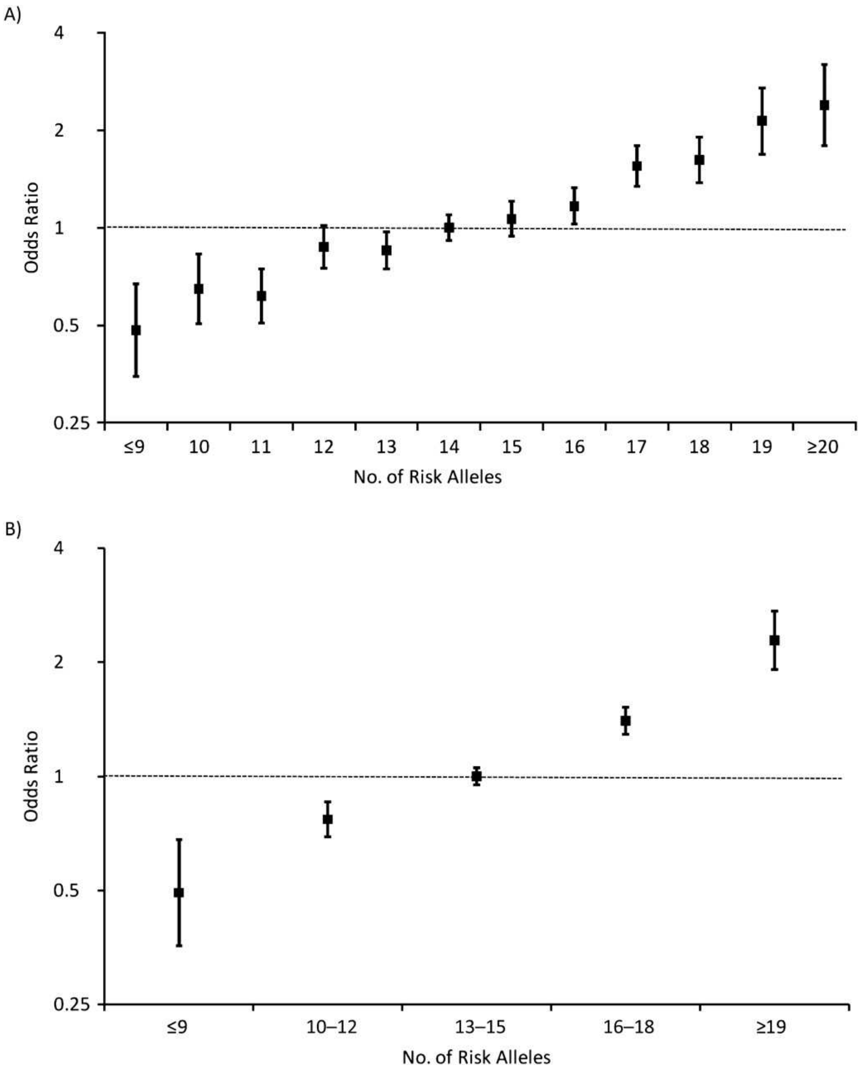


Figure 1.4 Plots showing the increasing odds ratios for colorectal cancer with the increasing number of risk alleles, London, United Kingdom, 1999–2007. The vertical bars represent 95% confidence intervals. The horizontal line denotes the null value (odds ratio = 1.0. Adapted from Lubbe *et al*, 2012.

1.4 Functional effects of low-penetrance CRC risk variants

As alluded to in the previous section, GWAS in general have detected risk variants with only modest effect sizes that are deemed too small to be meaningful. However, individually small effect sizes represent the reality of common genetic variation and do not necessarily preclude clinical utility. For instance, a GWAS hit for circulating lipid levels maps to the HMG-CoA reductase (HMGCR) gene (Kathiresan *et al*, 2008), the rate-limiting enzyme in cholesterol biosynthesis and the target of the extremely successful cholesterol-lowering statin drug. This discrepancy occurs because a drug's efficacy bears little relation to the degree of genetic variation in its target gene. Similarly, the size of the biological effect cannot be predicted by the epidemiological risk, or vice versa, not least due to pathway redundancies. Unravelling the mechanisms underlying GWAS associations will ultimately bring us closer to elucidating the genetic basis of complex disease, which in turn could identify novel causative biological pathways that may be suitable targets for chemopreventative drug development or repositioning of known therapeutics, as well as offer opportunities for personalised medicine.

The GWAS association signals in CRC have yet to be translated into a full understanding of the genetic elements that are mediating the effects of these susceptibility loci. Modern GWAS genotyping chips typically target SNPs chosen to capture variation across large genomic regions. These SNPs are not selected for having likely functional consequences, hence, hits from a successful GWAS merely mark a locus that encompasses one or more genetic variants that have biological functions driving the observed association with the trait phenotype. Contrary to early expectations, none of the GWAS-identified CRC risk variants are in protein-coding regions (Table 1.3). Assuming that the same is true for the candidate causal SNPs within the tagged haplotype block, the common heritability of CRC risk is thought to be mediated through genetic variation that influence gene regulation rather than protein sequence. The major challenge post-GWAS is to find the strongest candidate causal variants and identifying their target genes.

SNP	Locus	Closest gene	Relative position	Reference
rs10911251	1q25.3	*LAMC1*	Intronic	Peters *et al*, 2013
rs6687758	1q41	*DUSP10*	125Kb upstream	Houlston *et al*, 2010
rs6691170	1q41	*DUSP10*	250Kb upstream	Houlston *et al*, 2010
rs10936599	3q26.2	*MYNN*	Intronic	Houlston *et al*, 2010
rs1321311	6p21.2	*CDKN1A*	21Kb upstream	Dunlop *et al*, 2012
rs16892766	8q23.3	*EIF3H*	24Kb downstream	Tomlinson *et al*, 2008
rs6983267	8q24.21	*POU5F1B*	13Kb upstream	Zanke *et al*, 2007; Tomlinson *et al*, 2007
rs1035209	10q24.2	*SLC25A28*	25Kb downstream	Whiffin *et al*, 2014
rs10795668	10p14	*BC031880**	400Mb downstream	Tomlinson *et al*, 2008
rs3802842	11q23.1	*COLCA2*	Intronic	Tenesa *et al*, 2008
rs3824999	11q13.4	*POLD3*	Intronic	Dunlop *et al*, 2012
rs3217810	12p13.32	*CCND2*	Intronic	Peters *et al*, 2013
rs11169552	12q13.13	*DIP2B*	2.5Kb upstream	Houlston *et al*, 2010
rs7136702	12q13.13	*LARP4*	6.5Kb downstream	Houlston *et al*, 2010
rs1957636	14q22.2	*BMP4*	135Kb upstream	Tomlinson *et al*, 2011
rs4444235	14q22.2	*BMP4*	5.5Kb downstream	Houlston *et al*, 2008
rs11632715	15q13.3	*GREM1*	6Kb upstream	Tomlinson *et al*, 2011
rs16969681	15q13.3	*SCG5*	59Kb downstream	Tomlinson *et al*, 2011
rs9929218	16q22.1	*CDH1*	Intronic	Houlston *et al*, 2008
rs4939827	18q21.1	*SMAD7*	Intronic	Broderick *et al*, 2007
rs10411210	19q13.11	*RHPN2*	Intronic	Houlston *et al*, 2008
rs4813802	20p12.3	*BMP2*	49Kb upstream	Tomlinson *et al*, 2011
rs961253	20p12.3	*BMP2*	344Kb upstream	Houlston *et al*, 2008
rs4925386	20q13.33	*LAMA5*	Intronic	Houlston *et al*, 2010
rs5934683	Xp22.2	*GPR143*	Intronic	Dunlop *et al*, 2012

Table 1.3 The location of the tagging SNPs associated with CRC risk in relation to the closest gene. * denotes a predicted gene when there is no known protein-coding transcript in the vicinity

Identification of the truly causal variant requires a complete catalogue of all variants within the locus and the generation of such a catalogue has been the rate-limiting step. Fine-mapping of CRC risk loci is very much in its infancy, with most studies attempting to only narrow down the location of putative functional variants by imputation and targeted re-sequencing methods (Pittman *et al*, 2008; Carvajal-Carmona *et al*, 2011; Whiffin *et al*, 2013). While these studies suggest candidate regions, very few functional studies have been carried out to test these postulations.

The most well-studied CRC locus is the 8q24.21 locus, which despite its location in a gene desert, has pleiotropic effects on cancer susceptibility. Apart from its association with CRC, this locus also habours risk loci for solid tumours such as breast (Easton *et al*, 2007), prostate (Al Olama *et al*, 2009), ovarian (Ghoussaini *et al*, 2008) and bladder cancer (Kiemeney *et al*, 2008), as well as chronic lymphocytic leukemia (Crowther-Swanepoel *et al*, 2010). The rs6983267 SNP which is associated with increased risk of both colorectal and prostate cancers lie within an evolutionarily conserved region, and has been shown via computational predictions, enhancer reporter assays, chromatin-immunoprecipitation (ChIP) and transgenic mouse embryos to possess *in silico*, *in vitro* and *in vivo* properties of an enhancer, with allele-specific differential binding to the Wnt-regulated transcription factors TCF4 (Tuupanen *et al*, 2009; Sotelo *et al*, 2010) and TCF7L2 (Pomerantz *et al*, 2009). The target gene of this proposed enhancer element is not immediately obvious; although the well-known CRC proto-oncogene *MYC* lies ~335kb telomeric to rs6983267, there is a lack of association between the rs6983267 and gene expression in normal colorectal tissue or paired tumours. However, chromosome conformation capture (3C) techniques demonstrated long-range physical interaction between the enhancer element and *MYC* in a tissue-specific manner (Pomerantz *et al*, 2009; Sotelo *et al*, 2010). Altogether, the evidence from these functional studies suggests that the 8q24 risk locus acts as part of a cis-regulatory enhancer element for the *MYC* proto-oncogene, mediating CRC risk through its differential binding with TCF transcription factors.

Evidence for functionality at the 8q23.3 and 18q21 CRC risk loci have also been demonstrated by targeted re-sequencing and functional assays. At the 8q23.3 locus, the putative causal variant rs16888589 was validated using reporter gene assays and electrophoretic mobility shift assays (EMSA), with 3C analysis demonstrating a physical interaction between the encompassing control element and the promoter region of *EIF3H* located 144kb telomeric to the SNP (Pittman *et al*, 2010). At the 18q21 locus, a transgenic *Xenopus* model system was utilised to demonstrate that the putative causal variant intronic to *SMAD7* (Novel 1) is associated with a reduced expression of *SMAD7* in the colorectum (Pittman *et al*, 2009). There is also evidence that the tagging SNP rs4929827 is associated with *SMAD7* expression in human lymphoblastoid cell lines (Broderick *et al*, 2007).

Apart from these three loci described above, direct evidence implicating functionality of the remaining GWAS risk loci was scarce prior to the conception of this PhD project. Two *cis*-expression quantitative trait loci studies have since replicated part of my findings (Loo *et al*, 2012; Closa *et al*, 2014) and will be discussed whenever relevant in result Chapter 6.

1.5 eQTL

Landmark studies have clearly demonstrated that there is extensive natural variation in human gene expression within the same cell type and development stage, and that the gene expression phenotype is highly influenced by inherited DNA sequence variation (Cheung *et al* 2003; Morley *et al*, 2004; Stranger *et al*, 2005; Goring *et al*, 2007; Dixon *et al*, 2007). These non-coding germline variants are termed expression quantitative trait loci (eQTLs); they are referred to as local or *cis*-eQTLs when they map to the approximate location of the target gene, whereas those that map to considerable distances from the gene they regulate, often on different non-homologous chromosomes, are referred to as distant or *trans*-eQTLs. As the terminology *cis*- and *trans*- connote mechanism, it has been cautioned that this designation is best-reserved for use only when the functional variant has been identified (Rockman *et al*, 2006). The distinction between local and distant is arbitrary, and is usually pre-defined by study authors to be within 1-2 Mb of the variant under consideration.

1.5.1 Utility of eQTL in complex disease traits

eQTLs have been implicated in the predisposition to complex diseases in twin studies (Grundberg *et al*, 2012) as well as empirical studies of lymphoblastoid cell lines (Nicolae *et al*, 2010). By mapping the genetic architecture of gene expression in human tissues, eQTL studies can be useful in discovering candidate susceptibility genes for multifactorial diseases. The value of this has been illustrated by several proof-of-principle studies. Genome-wide transcriptional profiles of lymphocytes from the San Antonio Family Heart Study demonstrated that *cis*-eQTLs can be used as a discovery tool to identify novel candidate genes (and variants therein) that influence complex traits, e.g. *VNN1* gene and high-density lipoprotein cholesterol concentration (Goring *et al*, 2007). Another study examining genetic markers of childhood asthma incorporated eQTL analysis of EBV-transformed lymphoblastoid cell lines (LCL) as a component of the GWAS design, and utilised it to identify a novel candidate susceptibility gene *ORMDL3* for childhood asthma (Moffat *et al,* 2007). This has since spurred functional studies and transgenic mouse models demonstrating a role for this gene in asthma pathogenesis, providing valuable insights into the molecular mechanisms of proinflammatory diseases (Cantero-Recasens *et al*, 2009; Ha *et al*, 2013; Miller *et al*, 2014). Such findings have encouraged the use of eQTL data as a tool for interpreting results from GWASs, bridging the gap between common genetic markers for disease and the underlying mechanisms for clinical phenotypes. Importantly, eQTL annotation is carried out in an unbiased fashion, where the mapping of associations between alleles and target genes require no prior knowledge of functional mechanisms. Analyses of Crohn's disease are an example of this approach, where the biological effects of genetic markers were not readily deducible. Examination of LCL eQTL databases showed that one or more of these polymorphisms act as *cis*-acting factors influencing expression of genes (Libioulle *et al*, 2007). Similarly, the causal variant and causal gene for an LDL-cholesterol locus was identified by examining *cis*-gene expression levels in the liver and adipose tissue, providing the impetus for functional investigations of the implicated gene *SORT1* and its role in a novel lipoprotein regulatory pathway (Musunuru *et al*, 2010). More complex analysis of genetic variants that perturb networks through eQTL effects has provided important novel insights into the unravelling of complex trait genetics (Emilsson *et al*, 2008), as the genome exerts it functions through complex networks and multiple pathways that produce a wide range of responses. Hence, eQTL studies can be a powerful interpretive biological tool, and are integral in the systematic identification of transcriptional modules and construction of regulatory networks.

1.5.2 eQTL in the functional annotation of CRC risk loci

For several of the colorectal cancer risk loci, there is indirect evidence of an association with gene expression, but the evidence is circumstantial at best as they are largely based on SNPs in linkage disequilibrium with that tagging SNP. For example, the *CDH1* intron variant rs9929218 is in strong linkage disequilibrium with a *CDH1* promoter variant (Houlston *et al*, 2008), which has been reported to influence *CDH1* transcription in prostate cancer cell lines (Li *et al*, 2000). The 12q13.13 variant is in linkage disequilibrium with SNPs associated with *DIP2B* expression in lymphoblastoid cell lines, and the *LAMA5* intronic variant rs4925386 is in linkage disequilibrium with an eQTL for *LAMA5* expression in the liver (Houlston *et al*, 2010). However, the lack of an apparent effect on expression may merely reflect tissue-specificity of regulatory mechanisms. Most expression quantitative trait loci (eQTL) data sets are derived from only a limited number of source cell types, including monocytes, lymphoblastoid cells, liver and brain cells, and have not been comprehensively catalogued in colorectal tissue. This is a particularly important consideration as an estimated 50%-90% of eQTL are tissue dependent (Dimas *et al*, 2009; Nica *et al*, 2011), and trait-associated variants tend to exert more cell-type specificity (Fu *et al*, 2012; Brown *et al*, 2013). The other crucial aspect of the selection of tissue type in the measurement of eQTLs is the normality of the target tissue. Given that somatic alterations present in cancer cells can greatly affect expression (Figure 1.5), subtle genomic influences on expression can be masked (Curtis *et al*, 2012) and consequently be undetectable. Hence, it has been suggested that eQTL studies should be performed on non-aberrant cells representative of the cell of origin for the disease under study (Edwards *et al*, 2013).

Finally, it should be noted that the identification of an eQTL provides only an associative link between genotype and gene transcription; although it may imply causality, functional molecular approaches will be necessary to elucidate the underlying mechanism. Critically, even if a transcript is associated with a risk allele, this is not definitive of causation and functional follow up with assays relevant to the disease trait will be needed to demonstrate that a gene is directly involved in disease development.

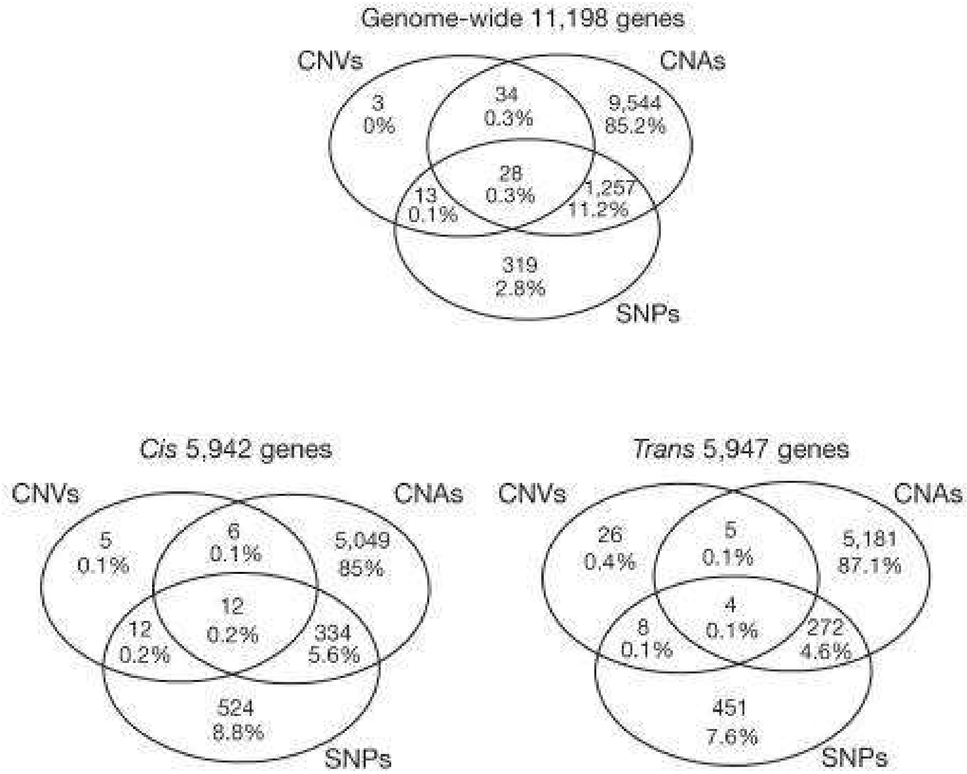


Figure 1.5 Somatic variants influence breast tumour expression architecture to a much greater extent than germline variants. Venn diagrams depict the relative contribution of SNPs, copy number variations (CNVs) and somatic copy number aberrations (CNAs) to genome-wide, *cis-* and *trans-* tumour expression variation for significant expression associations. Adapted from Curtis *et al*, 2012.

1.6 Research Aims

The large knowledge gap between the epidemiology and functional biology of common genetic variation in colorectal cancer, calls for studies that will translate CRC genetic associations into function. The overarching aim of this project is to functionally characterise these germline risk variants, with a view to improve the understanding of the biological mechanism(s) underlying these risk loci.

As all of the established risk loci reside within non-coding regions, it was hypothesised that they influence tissue-specific regulatory mechanisms and consequently, exert subtle effects on gene expression levels. The starting point for this project was to systematically investigate the functionality of common, low-penetrance risk variants using an unbiased eQTL approach that mirrors the agnostic style of GWAS. In view of the overall lack of direct association between these germline risk variants and expression in extra-colonic tissue types, it was hypothesised that any functional effects will be most prominent in the normal colorectal tissue, in particular the mucosal epithelial layer that harbours the cell of origin of colorectal cancer. Matched peripheral blood mononuclear cells (PBMC) will also be examined, as not only does this offer insight into the tissue-specificity of the underlying functional biology, any overlap between the two tissue-types could be advantageous in the context of identifying clinical biomarkers that are more easily accessible from patients. By integrating data from high-density DNA arrays and case-control series, the project also aims to identify the causal variant(s) that is most associated with specific gene expression as well as clinical risk.

Findings from the initial screening phases of the project will be rationalised and prioritised for further functional follow-up studies using molecular approaches. This is important, as few genes implicated in GWAS were previously evaluated in candidate gene studies. Surprisingly also, none of the currently identified loci are known to be involved in DNA repair, the principal pathway underscoring high-penetrance CRC susceptibility and a large proportion of sporadic CRC. Hence, evidence of the functional mechanism underlying these associations will not only provide support for the GWAS approach in the discovery of common risk variants, it will also allow the identification of target genes, offering new insights into the aetiology and pathogenesis of sporadic colorectal cancer.

1.7 Experimental approach

Several methodological approaches will be utilised in this project to bridge the gap between genetic risk and biological function, demonstrating a collaborative framework between clinicians, genetic epidemiologists and molecular biologists.

The principal theme of this project is to examine the association between genetic risk variants and gene expression in relevant non-aberrant tissue-types. To achieve this, normal mucosa specimens and matching blood from patients undergoing large bowel surgery were systematically collected and analysed. To simultaneously examine the expression of multiple genes, transcriptome-wide gene expression profiling with microarrays was utilised to maximise cost-effectiveness. As degradation of RNA compromises the ability to detect differential expression of genes especially those expressed at low levels, it is paramount that good quality, intact RNA was used to avoid poor data that may lead to erroneous conclusions. The isolation of intact RNA from the large bowel mucosa has inherent technical challenges and Chapter 3 focuses on the optimisation of this process to minimise RNA degradation that will have cascading detrimental effects on downstream experiments and analyses.

qRT-PCR technique is also a mainstay in this project due to the recurring themes of gene expression changes association with inherited variation. This highly-sensitive technique is of immense value in the same-sample validation of subtle differential gene expression results derived from microarray data, but its sensitivity is a double-edged sword and may lead to misleading results if not rigorously performed. The appropriate use of reference genes can vastly influence the accuracy of qRT-PCR results, and Chapter 4 concentrates on the selection and validation of stably expressed reference genes used to normalise the expression of genes of interest. Throughout Chapter 5, 6, 7 and 8, qRT-PCR quantification of gene expression will feature prominently in the assessment of differential expression and genotype-dependent functional differences, and the work presented in Chapter 4 allows confidence in the robustness of the data.

Prior to the analysis of gene expression in relation to genetic risk variants, expression was first evaluated in relation to clinico-pathological features in Chapter 5. This serves as a form of internal validation, and also highlights the need to adjust for such factors in the statistical analyses of the association between risk variants and gene expression.

Chapter 6 addresses the *cis*-eQTL analysis of the 25 CRC risk loci in normal large bowel mucosa and matching PBMC. This empirical approach is complemented by genotyping data from high-density arrays and imputation methods that aim to discover the functional variants underlying eQTL associations of risk loci.

The best eQTL associations are prioritised for technical validation, and Chapter 7 addresses the molecular mechanism underlying the Xp22.2 locus. This was performed by first using targeted re-sequencing and fine-mapping to identify putative causal variants of the *cis*-eQTL association. Thereafter, candidate causal variants and the tagging SNP were compared with interrogation of publically available functional data, reporter gene assays, transient siRNA knockdown approaches, and CRC case-control series. These observational and experimental approaches culminates in the identification of the causal variant at the Xp22.2 locus that best explains the association with the target gene *SHROOM2* as well as colorectal cancer genetic risk.

By inference, the target genes of the eQTL associations are likely to be associated with the risk of developing colorectal cancer. Chapter 8 describes functional follow-up assays to dissect the role and expression pattern of *SHROOM2*, a gene that has not previously been implicated in CRC. Functional phenotypes such as cell population doubling, wound closure and transcriptomic profiles were assessed after transient siRNA knockdown in cell lines, and less conventional localisation approaches were sought as the lack of a specific antibody precluded the utility of immuno-staining techniques.

As revealed in Chapter 6, genetic variation exerts their effect on CRC risk in part by influencing the expression of *cis*-genes. However, it is also evident that there are still a large proportion of risk variants whose functions and target genes cannot be accounted for by *cis*-eQTL effects on baseline expression. Chapter 9 demonstrates an alternative approach that mirrors that of epidemiological gene-environment studies to specifically investigate the 16q22.1 risk locus tagged by rs9929218, using the expression of the encompassing gene *CDH1* as the measure of outcome. Vitamin D levels and its pathway activity are postulated to modify the influence of rs9929218 on *CDH1* expression, and this hypothesis is tested in the normal large bowel mucosa, cell lines and human colonic epithelial crypt organoids.

Although detailed discussion of the results is provided in each result chapter, Chapter 10 summaries the main themes and conclusions that have emerged during the entire course of this research.

Overall, the work presented in this thesis demonstrates a multi-disciplinary approach in understanding the mechanisms underlying CRC GWAS-identified risk variants. It demonstrates that unbiased empirical approaches can be used to prioritise candidate variants/genes for follow-up functional studies, as well as the value of candidate gene/pathway approaches towards the ultimate goal of understanding the functional basis of CRC genetic predisposition.

Chapter 2

Materials and methods

This chapter describes the methods used in this thesis. More detailed methods are included in the results chapters where relevant. Where manufacturer's protocols have been used, these have been referred to, and any adjustments made to the cited method have been detailed in the text. Standard safety procedures and COSHH regulations were adhered to. All cell line culture were performed in a class 1 biological safety cabinet, whereas all biological material from primary tissue was handled in a class 2 safety cabinet. Reagents marked with an asterisk (*) were prepared by the technical services department at the MRC Human Genetics Unit, IGMM. Where the pH of solutions was adjusted this was done by adding concentrated HCl or NaOH as appropriate and monitoring of pH using a microprocessor pH meter (Hanna Interments).

2.1 Biological material

This study was set up and performed in collaboration with NHS Lothian/South East Scotland SAHSC Bioresource. All patients gave written informed consent. All information pertaining to subjects were in compliance with UK legislation and conform to the Tissue Act Scotland, 2006.

115 patients undergoing bowel resection operations for cancer, adenomas or non-malignant disease at the Western General Hospital, Edinburgh, were recruited for colonic tissue and peripheral blood sampling (detailed in Chapter 5). A further 40 patients undergoing bowel resection for colorectal cancer only were recruited for serial peripheral blood sampling (detailed in Chapter 9). Patients with known familial cancer syndromes, inflammatory bowel disease or those who have received pre-operative adjuvant chemo-radiotherapy were excluded. Recruitment and tissue/blood sampling was carried out over the course of this PhD with assistance from the group research nurse and technical staff from the Edinburgh Experimental Cancer Medicine Centre.

2.1.1 Sampling of fresh large intestinal mucosa and tumour

The resected bowel specimens were transported fresh to the pathology laboratory at room temperature immediately after each surgical resection. Macroscopic examination and assessment of the margins were performed by a pathologist for all specimens. Only tissue that is surplus to diagnostic requirement was taken. Undiseased colonic mucosa layers were dissected and separated from the muscularis at the resection margin furthest away from the tumour. Corresponding tumour were sampled by a pathologist whenever available and deemed to not interfere with the diagnosis. The fresh tissue samples were then flash-frozen in a cooling bath of 100% ethanol and dry ice, or stabilised in RNAlater® (Applied Biosystems) according to the manufacturer's protocol. Samples were then stored at -80°C before further processing.

2.1.2 Sampling of peripheral blood

Peripheral venous blood was drawn from patients using standard phlebotomy procedures. Blood for genomic DNA was collected in EDTA tubes, whereas blood for RNA and plasma extraction was collected in Lithium-Heparin tubes. When serial blood samples were required, surplus blood from clinical biochemistry requests were collected whenever possible to minimise the number of phlebotomy procedures for the patients.

Peripheral blood mononuclear cells (PBMCs) and plasma were isolated from approximately 9mls of fresh blood with Ficoll-Paque Plus (GE healthcare) according to the manufacturer's protocol. PBMCs were processed immediately for RNA extraction, whereas plasma was stored at -80°C until further analysis.

2.2 Cell culture

Media, solutions and additives:

Freezing medium

10% Dimethylsulfoxide (DMSO) (Sigma) in foetal calf serum*

Tissue culture medium

Cell-line specific basal medium (Table 2.1) (Life Technologies)

10% v/v foetal calf serum (FCS)*

1% v/v Penicillin and streptomycin*

Cell-line specific additional supplements (Table 2.1)

Phosphate buffered saline (PBS)*

0.1M $NaH_2PO_4.H_2O$

0.1M $Na_2HPO_4.7H_2O$

pH7.4

Trypsin Versene (T/V)

50% v/v Trypsin*

50% v/v Versene*

Cell lines stored in the liquid nitrogen facility at the MRC Human Genetics Unit were retrieved by rapid thawing in warm water and resuspending in 5mls of the appropriate tissue culture medium (Table 2.1) and then fed as required. After 3-4 days in culture, culture supernatant was sent to technical services for mycoplasma testing with the MycoAlert™ Mycoplasma Detection Kit (Lonza) to ensure all cells used were mycoplasma-free. To retain the cell lines as renewable sources, at least $3x10^6$ cells were split from the main culture, centrifuged at 1200rpm and the cell pellet resuspended in 1ml freezing medium. The cells were cooled immediately and then sequentially frozen at -80°C and -140°C.

Cell line	Tissue of origin	Nature of cells	Basal medium	Additional supplements
CACO2	Colorectal cancer	Adherent	DMEM	
COLO320	Colorectal cancer	Adherent	DMEM	
DLD1	Colorectal cancer	Adherent	DMEM	
HCT116	Colorectal cancer	Adherent	DMEM	
HELA	Cervical cancer	Adherent	DMEM	
HEK293	Embryonic kidney	Adherent	DMEM	
HT29	Colorectal cancer	Adherent	DMEM	
K562	Erythroleukemia	Suspension	RPMI	
LOVO	Colorectal cancer	Adherent	DMEM	
MCF7	Breast adenocarcinoma	Adherent	DMEM	
PNT	Prostate epithelium	Adherent	DMEM	
RKO	Colorectal cancer	Adherent	DMEM	
RPE1	Retinal pigment epithelium	Adherent	DMEM/F12	1% v/v Glutamine*
SW48	Colorectal cancer	Adherent	DMEM	
SW480	Colorectal cancer	Adherent	DMEM	
VACO425	Colorectal cancer	Adherent	DMEM	

Table 2.1 List of cell lines used and their characteristics.

2.2.1 Maintenance of adherent cell lines

For adherent cell lines, media was changed every 3-4 days to maintain cells in the logarithm phase of growth. Cell lines were passaged at 80-90% confluence using T/V solution after washing the cells with warm PBS.

2.2.2 Maintenance of suspension cell lines

Suspension cell lines were grown in upright T-flasks and periodically shaken to break up the cell clumps. Cultures were fed every 2-4 days depending on the population doubling time, by removing half of the media from the flask and replacing it with a slightly increased volume of fresh media. Cultures were split when the cell count is approximately 2 x 10^6 cells/ml, with a minimum cell concentration of 200 x 10^3 cells/ml for each subculture to ensure optimal growth. Cells were counted with a Coulter Counter® (Beckman Coulter).

2.3 RNA work

2.3.1 RNA extraction

Cell lines

Adherent cells of 80-90% confluence were detached from T25 flasks with a cell scraper into 2mL of cold PBS. The cells were pelleted by centrifugation at 1600rpm, resuspended in 1ml of TRIzol (Life Technologies), and total RNA extracted according to the manufacturer's protocol. Alternatively, RNA from cell lines was extracted directly from the culture plates with the RNeasy Mini Kit (Qiagen) according to the manufacturer's instructions. Both techniques produced comparable RNA yield and quality.

Fresh frozen human large intestinal mucosa and tumours

The method for extracting total RNA from fresh frozen human large intestinal mucosa required optimisation. The method presented here is the final optimised method; the optimisation process will be discussed in Chapter 3. Fresh frozen human large intestinal samples (no larger than 0.5cm in the smallest dimension) were transitioned to RNAlater-ICE® (Applied Biosystems) on dry ice and kept at -20°C for 16 hours before storage at -80°C according the manufacturer's protocol. This was not necessary for samples already stabilised in RNAlater (Applied Biosystems) upon collection. The TissueLyser LT (Qiagen) and a single 0.2mm stainless steel bead were used for the mechanical disruption and homogenisation of samples. Total RNA was then isolated using the RiboPure Kit (Applied Biosystems) according to the manufacturer's instructions.

Fresh PBMC

RNA extraction for PBMCs were performed immediately after isolation from whole venous blood, using the Ambion® RiboPure™ Kit (Applied Biosystems) according to the manufacturer's instructions.

All RNA was stored at -80°C until further analysis.

2.3.2 Evaluation of RNA quality and yield

Gene expression profiling microarrays

RNA purity was measured using the Nanodrop® 800 spectrophotometer (Thermo Scientific), and all samples used for gene-expression profiling had the ratio of absorbance at 260nm and 280nm (A260/A280) of >1.8. RNA yield and integrity was measured using the 2100 Bioanalyzer® (Agilent Technologies). The RNA integrity numbers were ≥7 for normal mucosa RNA samples, ≥8 for PBMC RNA samples, and ≥9 for cell line RNA samples.

PCR

For downstream experiments with RT-PCR and qRT-PCR, RNA yield and purity was measured with the Nanodrop® 800 spectrophotometer (Thermo Scientific), with all samples giving A260/A280 ratios of >1.8.

2.3.3 DNase treatment of RNA

RNA samples were first treated with DNase in 10µl reactions. Final reaction concentrations were 100ng/µl total RNA, 1x RQ1 RNase-free DNase reaction buffer (Promega), 0.1 unit/µl RQ1 RNase-free DNase (Promega). The reaction was incubated at 37°C for 30 minutes. 1µl of RQ1 DNase Stop Solution (Promega) was then added and the reaction incubated at 65°C for 10 minutes to inactivate the DNase.

2.3.4 cDNA synthesis from RNA

The DNase treated RNA samples were reversed transcribed to cDNA using Moloney Murine Leukemia Virus Reverse Transcriptase (M-MLV RT) (Promega) in a 20µl final reaction volume. The final reaction concentrations were:

50ng/µl input RNA

10units/µl M-MLV RT (Promega)

1x M-MLV RT reaction buffer (Promega)

0.4units/µl random primers (Roche)

1mM dNTP (Promega)

1 unit/µl RNasin® Ribonuclease Inhibitor (Promega)

2.3.5 Quantitative real time PCR (qRT-PCR)

cDNA from cell lines were diluted to 1:20 working stock whereas cDNA from patient samples were diluted to 1:10 working stock. qRT-PCR was carried out in 10µl final reaction volumes:

2µl cDNA working stock

5µl Taqman® Master Mix (Applied Biosystems)

0.5µl Taqman® Gene Expression Assay (Applied Biosystems)

2.5µl nuclease-free H2O

The linearity and amplification efficiency of each individual gene expression assay were first tested using serial dilutions of cDNA from an expressing cell line to produce calibration curves. The threshold cycles (Ct) were plotted against the logged cDNA quantity, and the coefficient of correlation obtained for the fitted calibration curves (R^2) were calculated. PCR amplification efficiencies were calculated from the slope of the log-linear portion of the calibration curves using the equation

$$\textit{Efficiency} = 10^{(-1/slope)} - 1$$

All assays used had highly linear calibration curves (R^2 of >0.985) and efficiencies between 90% - 110%, indicating that the assays are well-optimised and the input template is of good quality.

All reactions were performed in triplicates. Amplification and detection of the amplified product was carried out with ABI PRISM HT 7900 Sequence Detection

System thermal cycler (Applied Biosystems) and read using SDS Version 2.3 software (Applied Biosystems). The PCR reaction conditions were: 50°C – 2min, 95°C – 10min, 95°C – 15secs, 60°C – 1min. The cycle was repeated 40x.

For detection of differential expression in the normal colorectal mucosa, the expression of genes of interest was normalised using three validated reference genes. The selection and validation of reference genes are described and discussed in more detail in Chapter 4. In the PBMCs, the housekeeping gene *GAPDH* was used as a reference gene for the normalisation of expression quantification.

2.3.6 Whole transcriptome gene expression profiling

Expression profiling was accomplished using the HumanHT-12 v4.0 Expression BeadChip Arrays (Illumina, USA). Each array contains 50-mer probes representing more than 47,000 transcripts derived from the NCBI RefSeq Release 38 (November 7, 2009),[1] as well as legacy UniGene content. RNA was amplified and biotin-labelled using Ambion's Illumina Total Prep RNA Amplification Kit (Ambion), as per manufacturer's protocol. 500ng input total RNA was used, and *in vitro* transcription incubation was carried out for 14 hours. The quality and yield of the amplified RNA (aRNA) was assessed with the 2100 Bioanalyzer® (Agilent Technologies) to ensure that the aRNA profile is as expected, producing a distribution of sizes from 250 to 5500 nt with most of the aRNA at 1000 to 1500nt. This was then sent to Genetics Core, Wellcome Trust Clinical Research Facility, Edinburgh, for array hybridization and scanning.

[1] ftp://ftp.ncbi.nih.gov/refseq/release/release-notes/archive/RefSeq-release38.txt

2.4 DNA work

2.4.1 DNA extraction from whole blood

Isolation of genomic DNA from whole blood was carried out using a Nucleon™ BACC Genomic DNA Extraction Kit (GE healthcare) according to the manufacturer's instructions. DNA was suspended in TE and quantified using the Nanodrop 800 spectrophotometer (Thermo Scientific).

2.4.2 PCR

PCR reactions were performed in a final volume of 25µl using platinum Taq® DNA polymerase (Invitrogen). Final reaction concentrations were 1x PCR buffer, 0.2mM dNTPs, 1uM oligonucleotide primer (forward and reverse), 5ng DNA, 2.5mM magnesium chloride, and 1 unit of platinum Taq® DNA polymerase. All reagents used were supplied by Invitrogen. Amplification was performed using a Peltier PCT225 thermal cycler (MJ Research) under the following standard conditions: 95°C for 5 minutes, (95°C for 30 seconds, 62°C for 30 seconds, 72°C for 30 seconds) x 30 cycles, 72°C for 5 minutes.

Primers were supplied by Sigma as precipitates and re-suspended in dH2O to a stock concentration of 20uM. These oligonucleotides will be referenced in the relevant chapters.

2.4.3 Gel electrophoresis

Solutions:

<u>10 x Tris-Acetate EDTA (TAE)*</u>

2M Tris

5.7% w/v Glacial acetic acid

50mM Na_2EDTA (pH8.0)

Loading buffer

100mM Na_2EDTA (pH8.0)

0.25% w/v Bromophenol blue

30% w/v Sucrose

All PCR products were visualised on a gel before sequencing. 2% agarose gels were prepared using routine electrophoresis grade agarose (Biogene Ltd) and 1x TAE electrophoresis buffer. 5µl of ethidium bromide (BioRad) per 100ml of gel mixture was added to the gel for visualisation.

5µl of PCR products were loaded onto the gel with 3µl of loading buffer. Size markers used were generally a 1kb ladder (Invitrogen). The DNA was electrophoresed at 40-60V and visualised by UV trans-illumination using a Herolab trans-illuminator (Herolab, Weisloch). Images were visualised on the BioRad Chemidoc system using QuantityOne software (Biorad).

2.4.4 Purification of PCR products

PCR products were first purified with the Exo-SAP clean-up protocol. 7.5µl reactions were prepared with:

3µl of PCR product

3.75µl of dH2O

0.25µl of Exonuclease I (10units/µl) (USB)

0.5µl of Shrimp Alkaline Phosphatase (1unit/µl) (USB)

Reactions were incubated under the following conditions using a Peltier PCT225 thermal cycler (MJ Research) at 37°C for 15 minutes, 80°C for 15 minutes, and 4°C for 10 minutes.

2.4.5 DNA sequencing

Sequencing of the purified PCR products were carried out in 10µl reactions using either individual Eppendorf tubes or in 96-well plates. The reactions consists of:

3.5µl of purified PCR product

5µl of dH2O

1µl of Big Dye® Terminator V3.1 Kit *

1µl of 20uM oligonucleotide primer (forward or reverse)

Amplification was performed on a Peltier PCT225 thermal cycler (MJ Research) at 96°C for 30 seconds, 50°C for 15 seconds, 60°C for 4 minutes, for 25 cycles.

2.4.6 Precipitation of DNA from sequencing reactions

After amplification, precipitation of sequenced DNA was carried out by adding 50µl 95% ethanol and 2µl 3M NaOAC (pH 4) to each sequencing reaction mix. The mixture was incubated for 30 minutes at room temperature and then centrifuged at 1200rpm for 30 minutes. The majority of the supernatant was removed from the wells by inverting the plates or tubes, and the residual removed by pulse spinning the upturned plates/tubes on paper towels at 800rpm. The DNA pellets were washed by adding 70% ethanol down the side of the wells and inverting the plate immediately to remove the supernatant. The pellets were dried by a further pulse spin and stored at -20°C. The precipitated reaction products were re-suspended in HiDi^{TM} (Applied Biosystems), heated at 90°C for 2min, and resolved on the ABI PRISM$^{®}$ 3100 or 3730 genetic analysers by Technical Services, MRC Human Genetics Unit, IGMM.

2.4.7 Analysis of sequence data

Sequence data was analysed using Consed (Gordon *et al*, 1998) and Mutation Surveyor V3.30 (Biogene).

2.5 Protein biology

Solutions:

Lysis Buffer

100µl of 10X Whole Cell Lysis Buffer (Cell Signalling Technology)

40µl of 25X Complete™ Protease inhibitor cocktail (Roche Diagnostics)

10µl of 100mg/ml Pefabloc SC (Roche Diagnostics)

1µl of 1mg/ml Pepstatin A (Sigma)

1µl of 1M NaF (Sigma)

1µl of 1M Na3VO4 (Sigma)

5µl of 200mM phenylmethanesulfonylfluoride (PMSF) (Sigma)

842µl of dH2O

6x Sample Buffer

20% w/v Glycerol

2% w/v Sodium dodecyl sulfate (SDS)

0.25% w/v Bromophenol blue

1x Stacking buffer

5% w/v β-mercaptoethanol

4x Resolving Buffer

1.5M Tris

0.4% w/v SDS

pH 8.8

4x Stacking Buffer

500mM Tris

0.4% w/v SDS

pH 6.8

10x Running Buffer

250mM Tris

2M Glycine

1% w/v SDS

10x Wet Transfer Buffer

250mM Tris

2M Glycine

1x Wet Transfer Buffer

10% v/v 10x Wet Transfer Buffer

10% v/v 100% methanol

5% Resolving Gel

1x Resolving buffer

5% w/v Acrylamide

0.15% w/v Ammonium persulphate (APS)

0.01% w/v N,N,N’,N’,tetramethyl-1-2-diaminomethane (TEMED)

8% Resolving Gel

1x Resolving buffer

8% w/v Acrylamide

0.15% w/v APS

0.01% w/v TEMED

<u>4% Stacking Gel</u>

1x Stacking buffer

4% w/v Acrylamide

0.15% w/v APS

0.01% w/v TEMED

2.5.1 Preparation of total protein extracts

Human normal colorectal mucosa

The TissueLyser LT (Qiagen) and a single 0.2mm stainless steel bead were used for the mechanical disruption and homogenisation of fresh frozen mucosa samples in lysis buffer. The lysis reaction was then incubated for 30 minutes on ice, with vortex mixing every 10 minutes. Debris was cleared by centrifugation at 13200rpm for 10 minutes at 4°C. The supernatant containing total protein extract was collected and stored at -40°C.

Cell lines

Adherent cells of 80-90% confluence were detached from T25 flasks with a cell scraper into 2mL of cold PBS. The cells were pelleted by centrifugation at 1600rpm and resuspended in 100µl of lysis buffer. Incubation, extraction and storage conditions were the same as the procedure for colorectal mucosa.

2.5.2 Cellular subfractionation of protein extracts

In order to quantify protein in the different subcellular compartments, DLD1 cells were grown to confluence in T75 flasks. The cells were then extracted using the ProteoExtract® Subcellular Proteome Extraction Kit (CalbioChem) as per the manufacturer's instructions, allowing subfractionation of cellular protein in the cytoplasmic, membrane/organelle, nuclear and cytoskeleton compartments. Briefly, this kit utilises four specialised extraction buffers to sequentially extract the different subcellular compartments based on the differential solubility of proteins in each compartment. A schematic overview is shown in Figure 2.1.

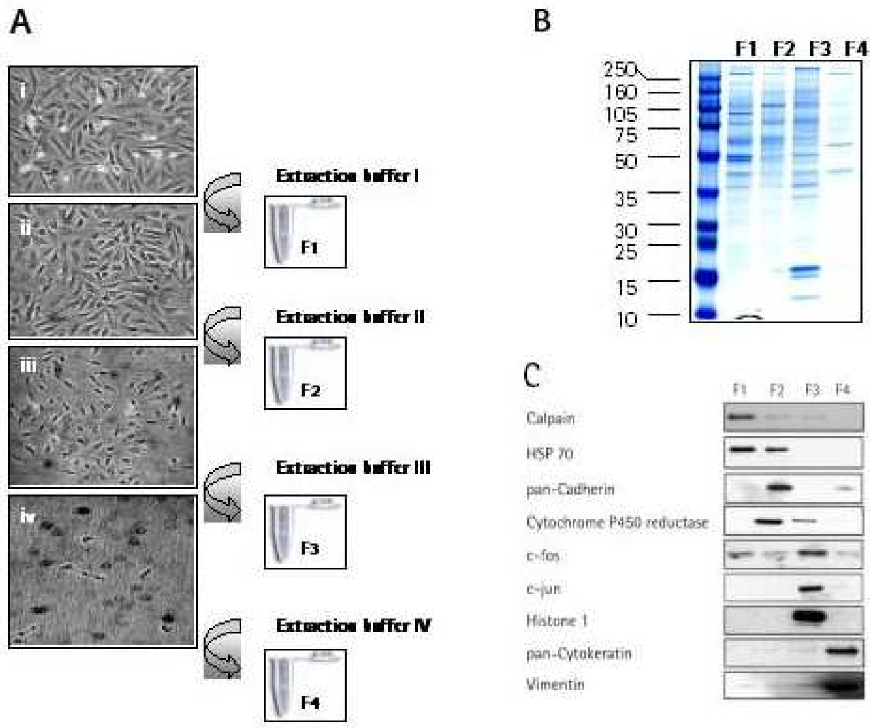


Figure 2.1 Illustration from the CalbioChem ProteoExtract® Kit protocol demonstrating the steps involved in the extraction of subcellular compartments. Four fractions are extracted enriched for: cytosolic fraction (F1), membrane/organelle protein fraction (F2), nucleic protein fraction (F3), cytoskeletal fraction (F4). A) Adherent SAOS cells were extracted using sequential buffers. Images show cells before and after the extraction with the respective extraction buffer. B) SDS-PAGE analysis of each subcellular fraction demonstrates distinct protein patterns of the respective fractions. C) The selectivity of subcellular extraction was demonstrated by immunoblotting against the indicated marker proteins.

All protein extracts were stored at -40°C. The concentrations of the total protein extracts were determined by Bradford assays (Biorad) using bovine serum albumin (BSA) (Sigma) to generate a standard curve. All samples were measured in triplicate and concentrations calculated from the concentration gradient of the BSA standard curve. The extracts were then diluted in water to the same concentrations for Western blot analysis.

2.5.3 Western Blot analysis

30-50µg of total protein extract was added to a 1:7.5 dilution of sample buffer and boiled for 3-5 minutes. Samples were resolved by denaturing SDS-PAGE on a 5%/8% polyacrylamide gradient gel in 1x running buffer. Alternatively, a precast 4-12% Bis-Tris Polyacrylamide Gel (NuPAGE® Novex®, Life Technologies) was used according to the manufacturer’s protocol. Pre-stained molecular weight markers (Biorad) were run in parallel. PVDF membranes (Biorad) were prepared by immersion in 100% methanol for 2 minutes. Protein was transferred from gels to membranes by wet transfer for 1 hour at 100V, at 4ºC, using 1x wet transfer buffer. This is followed by blocking of the blots in 5% w/v dried milk (Marvel) and 0.15% v/v Tween (Sigma) in PBS for 2 hours at room temperature. Blots were then probed with primary antibodies overnight at 4ºC in PBS/BSA/Azide*. Blots were washed afterwards in 0.15% Tween/PBS for 20 minutes x 3 with gentle shaking, then incubated in the appropriate species-specific horseradish peroxidase (HRP)-conjugated secondary antibody in 5% milk/0.15% Tween/PBS for 1 hour at room temperature. Blots were again washed as previously before detection of specifically bound antibody by chemiluminescence using Luminol reagent (Santa Cruz Biotechnology). Luminol reagent was applied to the blots for 1 minute, followed by covering the blots with a plastic cover and exposure to Amersham Hyperfilm™ ECL (GE Healthcare).

Primary antibodies and dilutions are detailed in the relevant chapters. Goat anti-rabbit IgG-HRP, goat anti-mouse IgG-HRP and donkey anti-goat IgG-HRP secondary antibodies (Santa Cruz) were used at 1:3000 dilutions.

2.6 Data analysis

2.6.1 Statistical analysis

Statistical analyses were performed using GraphPad, Excel and R. The tests used for each individual result are reported in the figure legends. Unless stated otherwise, *p*-values of <0.05 was the significant threshold for reporting. The single asterisk * indicates $p<0.05$, the double asterisk ** indicates $p<0.01$, and the triple asterisk *** indicates $p<0.001$.

2.6.2 Analysis of microarray gene expression data[2]

Microarray data, exported from Beadstudio, was processed and normalized using the R, Bioconductor beadarray and *limma* packages. Prior to normalization probes that were not detected (detection p-value > 0.01) on the microarrays were removed. Microarrays were quantile normalized to remove technical variation. The average signal of the biological replicates (n=3) were used for further analysis. ComBat batch correction was performed to control for batch effects. The *limma* package was used to find differential expressed genes, using the functions lmFit, eBayes and topTable. Unless stated otherwise, age, gender, presenting pathology (cancer vs non-cancer) and the anatomical sampling site were used as co-variates in the analyses as appropriate. Multiple testing correction was calculated using FDR *q*-values (Benjamini & Hochberg, 1995) to minimise false negatives.

2.6.3 Genomic annotations and functional predictions

The genome browsers Ensembl and UCSC browser (Kent et al, 2002; URL2.1) were used to interrogate publically available databases. All annotations were presented according to the human reference sequence build GRCh37.p12 (hg19). SIFT (Kumar et al, 2009) and PolyPhen (Adzhubei et al, 2010) were used for predicting the effects of coding non-synonymous variants on protein function. The JASPER matrix model (Mathelier et al, 2013) was utilised to predict the transcription factor binding

[2] Analysis performed by Graeme Grimes and Victoria Svinti, MRC Human Genetics Unit, IGMM.

affinities of oligonucleotide sequences. Functional enrichment analysis of differentially expressed genes was performed using the web-accessible tool DAVID v6.7 (Huang *et al*, 2009; URL2.2) and GOrilla (Eran *et al*, 2009; URL2.3).

Chapter 3

Isolation of high-quality intact RNA from human colorectal normal mucosa

3.1 Introduction

The extraction of intact RNA from fresh frozen human colonic normal mucosa has been a technically challenging aspect of this study. Various factors can have an undesirable impact on the RNA integrity, such as ischaemic time (Huang *et al*, 2001; Spruessel *et al*, 2004), endogenous RNases, exposure to environmental RNases and freeze-thawing during the processing of tissue samples (Botling *et al*, 2009). It is widely recognised that intact input mRNA is critical for gene expression array analysis, as using degraded mRNA may result in misleading variability and transcriptional differences. Conventionally, the 28S:18S rRNA ratio has been used as a measure of mRNA quality, with a 2:1 ratio considered the benchmark for intact RNA. However, this method is somewhat subjective because the appearance of the rRNA bands can be affected by the electrophoresis conditions, the amount of RNA loaded and the saturation of ethidium bromide fluorescence (Palmer *et al*, 2004). Moreover, relatively large amounts of RNA are required for the gel electrophoresis, which may not be possible when there is limited sample amount from small human biopsies. The Agilent 2100 Bioanalyzer is an increasingly used analytical tool for total RNA analysis, using a combination of microfluidics, capillary electrophoresis and fluorescence to evaluate concentration and integrity. The RIN (RNA integrity number) generated by an automated algorithm has been shown to be an effective method to assess RNA quality (Strand *et al*, 2007; Copois *et al*, 2007, Schroeder *et al*, 2006). A RIN of ≥7.0 is generally accepted as adequate integrity for amplification and microarray analysis of human tissues. This chapter will discuss some of the challenges I have faced and present the results of several RNA extraction techniques in an effort to establish a replicable and robust extraction protocol for microarray quality RNA, from fresh frozen human colorectal normal mucosa.

3.2 Methodological overview

3.2.1 Study subjects and biological material

To explore the functional effects of common genetic risk variants in the normal colonic mucosa, 115 fresh normal mucosa were harvested immediately after surgical resection of colorectal adenocarcinoma (n=99), tubulo-villous (n=8) and villous (n=2) adenomas or non-neoplastic conditions (n=6), as described in Chapter 2. The tissue samples collected were flash-frozen in ethanol/dry ice or allowed to equilibrate in the RNA-stabilising solution RNA*later*® (Life Technologies) to preserve RNA integrity.

3.2.2 RNA extraction

RNA extraction was performed using guanidinium thiocyanate-phenol-chloroform extraction methods with TRIzol® reagent (Life Technologies) and the Ambion® RiboPure™ Kit (Life Technologies), according to the manufacturers' protocols. To optimise the quality and quantity of RNA extracted, several technical aspects to the protocols were modified and will be discussed in more detail in 3.3. Specifically, these include the use of RNAse inhibitors (Superase•In™, Life Technologies), RNA*later*®-ICE Frozen Tissue Transition Solution (Life Technologies), and mechanical disruption of the mucosa samples with a bead mill homogeniser TissueLyser LT (Qiagen).

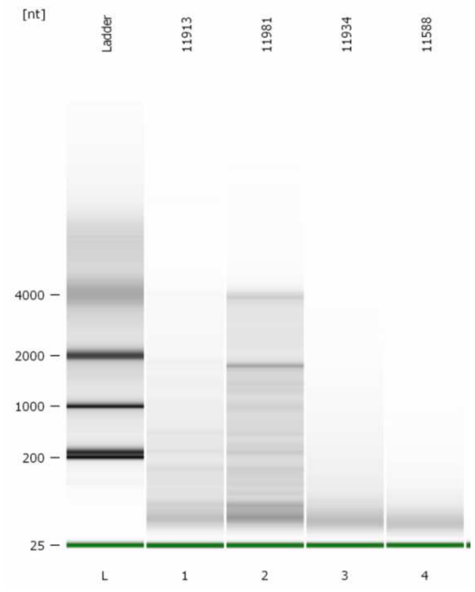
3.2.3 Evaluation of RNA quality and yield

RNA purity was evaluated using the Nanodrop® 800 spectrophotometer (Thermo Scientific), whereas RNA integrity and yield was measured with the 2100 Bioanalyzer® (Agilent Technologies).

3.3 Results

3.3.1 RNA-stabilising solutions and RNase inhibitors

The existing in-house protocol for RNA extraction from primary tissue utilises TRIzol® reagent (Life Technologies) using an adapted protocol. Flash-frozen tissue was cut out from cryovials and then manually disrupted with a scalpel before immersing in TRIzol®. This process of disruption was continued by grinding of the tissue with a mini-pestle. This was followed by phase separation, RNA precipitation, wash and redissolving as per the manufacturer's protocol. Although this procedure was performed on ice quickly, the yield of the extraction ranged from 620-2720ng, with RINs ranging from 2.1-4.2 (Figure 3.1), which were suboptimal for gene expression assays.



Sample	A260/A280	Yield (ng)	RIN
11913	1.94	1380	2.1
11981	1.94	2720	4.2
11934	1.92	980	2.2
11588	1.86	620	2.1

Figure 3.1. Quality parameters of the RNA extracted with the in-house RNA extraction protocol utilising Trizol® in four representative normal mucosa samples (lanes 1-4). Digital gel electrophoresis was performed using the 2100 Bioanalyzer® (Agilent). Normal mucosa RNA samples were run concurrently with the RNA 6000 Nano ladder (Agilent) containing six RNA fragments ranging in size from 0.2 to 6.0 kb.

To optimise the recovery of intact RNA, a trial of RNA-stabilising solutions and RNA inhibitors was carried out on samples collected from the same patient at the same time (Table 3.1). RNA*later*® is an aqueous solution designed specifically to preserve RNA integrity during the storage of fresh tissue and cells, thus circumventing the freeze-thawing process that compromises RNA integrity. RNA*later*®-ICE is designed by the manufacturer for use on samples that are already frozen, allowing tissue to be transitioned from a frozen to non-frozen state for processing. As I already had a collection of snap-frozen tissue, RNA*later*®-ICE was trialed alongside RNA*later*® to assess the comparability of the two solutions.

Sample	Treatment	Details of modification	Yield (ng)	A260/ A280	RIN
A	Trizol extraction	None	7840	1.93	2.9
B	RNA*later*®-ICE	Snap-frozen tissue sample (-80°C) immersed in RNAlater-ICE (-80°C) and allowed to thaw at -20°C for 16 hours.	8880	1.98	5.5
C	RNA*later*®	Tissue sample immersed in RNAlater immediately after sampling, equilibrated at 4°C for 16 hours before discarding the solution and storing tissue at -80°C.	1520	1.93	2.8
D	RNase Inhibitor	80U of Superase•In™ placed onto tissue during mechanical disruption and 40U added into the supernatant collected after phase separation.	8000	1.86	2.4
E	RNA*later*®-ICE + RNase Inhibitor	As per B and D	5560	1.87	5.1
F	RNA*later*® + RNase Inhibitor	As per C and D	4720	1.71	7.6

Table 3.1. Modifications to the RNA extraction protocol was performed on samples obtained from the same patient at the same time.

These initial results suggest that a combination of RNA*later*® and an RNase inhibitor would provide the most protection against RNA degradation during the extraction process. However, the effect of the RNase inihibitor is unclear, as there appears to be no additional protective effect on untreated samples and samples treated with RNA*later*®-ICE.

3.3.2 Bead mill homogenisation and glass-fibre filter RNA purification

As there were a significant number of tissue samples already snap-frozen on collection, I focused on using RNA*later*®-ICE and RNase inhibitors to stabilise these samples. Although there was an improvement in the RIN values of some of these samples, it was not consistently reproducible across all samples (Table 3.2). As RNA degradation can occur quickly within the first few minutes of the tissue thawing, it is likely that RNA degradation has already occurred within some of these tissue samples, when they were removed from -80ºC storage and handled for previous extractions. It is also possible that degradation occurred as a result of inefficient lysis and homogenisation.

	In-house protocol			**RNAlater®-ICE and Superase•In™**		
Sample	A260/280	Yield (ng/µl)	RIN	A260/280	Yield (ng/µl)	RIN
11913	1.94	1380	2.1	1.87	2180	6.2
11981	1.94	2720	4.2	1.87	4600	2.5
11934	1.92	980	2.2	1.88	6840	5.6
11588	1.86	620	2.1	1.81	1920	2.5

Table 3.2. Quality parameters of RNA when samples were re-extracted with modifications using RNAlater-ICE and RNase inhibitors.

To seek further improvement in the quality of isolated RNA, replicate samples that have not previously been removed from -80ºC storage were used, and several additional modifications to the protocol were made. A bead mill tissue homogeniser (TissueLyser LT, Qiagen) was used for more thorough mechanical disruption and homogenisation of the tissue, using a single 0.2mm stainless steel bead for a tissue biopsy of approximately half the size of the bead. The process of mechanical disruption was performed on dry ice instead of ice to prevent any thawing of the tissue samples. Additionally, a commercial RNA extraction kit Ambion®

RiboPure™ Kit (Life Technologies) that combines lysis with TRI Reagent® lysis and glass-fibre filter RNA purification was also used. The glass-fibre filter removes residual proteins and lipids as well as smaller degraded RNA fragments. The procedure is compatible with tissues that have been treated with either RNAlater® or RNA*later*®-ICE. There was a marked improvement in the isolated RNA (Table 3.3), which was consistently replicated in subsequent extractions of more samples. This optimised protocol also produced similarly intact RNA from RNA*later*® stablised samples (RIN>7). The RNAse-inhibitor was eventually removed from the protocol as it was not observed to further improve the quality of the extracted RNA.

	RNAlater®-ICE and Superase•In™			**TissueLyser LT and RiboPure™ kit**		
Sample	A260/280	Yield (ng/µl)	RIN	A260/280	Yield (ng/µl)	RIN
11913	1.87	2180	6.2	2.12	3620	9
11981	1.87	4600	2.5	2.17	2420	7.5
11934	1.88	6840	5.6	2.30	2050	8.8
11588	1.81	1920	2.5	2.28	3900	8.4

Table 3.3. Quality parameters of RNA when samples were re-extracted with modifications using to the protocol, using a bead mill homogeniser and glass-filter RNA purification.

3.4 Discussion

The extraction of RNA can be greatly complicated by the presence of ubiquitous and hardy ribonucleases that degrade RNA; these can be cellular RNases that are released from the cells, or those present in the environment. The isolation of RNA from normal colonic mucosa is particularly challenging as not only it is rich in RNAses, it is also much tougher compared to friable tumour tissue. The optimisation of the RNA extraction process has demonstrated to me the technical challenges of preserving the quality of RNA to ensure accuracy of downstream experiments and observations. This learning process highlights the importance of robust and replicable techniques, as well as principles of optimisation, experimental planning and the use of controls. Ultimately, it facilitated and enabled the reliable analysis of genome-wide expression for the 115 colorectal mucosa samples collected from surgical resections.

Chapter 4

Selection of reference genes for qRT-PCR quantification of gene expression

4.1 Introduction

Quantitative real time PCR (qRT-PCR) is an accurate, fast and sensitive measurement of gene expression. It is the method of choice to validate the results of microarray analysis, as well as to perform independent analysis on a defined number of genes on validation sets and cell lines. The sensitivity of this technique also means it is prone to confounding variation resulting from factors such as the quantity and quality of the template input, as well as the yields and efficiency of the extraction and the enzymatic reactions. A robust normalisation technique is therefore required to remove experimentally-induced non-biological variations and minimise quantification error.

The use of reference genes as internal controls is currently the preferred normalisation method (Huggett *et al*, 2005), but there is increasing evidence that the expression of commonly used reference genes are context dependent and can vary significantly between tissue types (Barber *et al*, 2005) and experimental conditions (Dheda *et al*, 2005). If unrecognised, expression changes in reference genes can lead to erroneous conclusions about real biological effects. This is a particularly important point to address in my study, as the differential expression of target genes associated with common genetic variation is likely to be subtle and may be easily masked by any changes in the reference genes.

It is now recommended that normalisation against three or more validated reference genes is the most appropriate and universally applicable method (as reviewed by Derveaux *et al*, 2010). To select reference genes for a sample set, a pilot study should be performed to measure 10 candidate genes in 10 representative samples. Using the raw non-normalised expression values, the expression stability can then be analysed using various mathematical algorithms and software e.g.

geNorm, BestKeeper and NormFinder (as reviewed by Vandesompele *et al*, 2009). The underlying principle of these algorithms is that the expression ratio of two proper reference genes should be constant across samples.

This chapter describes the results of the optimisation and selection of reference genes which would enable validation of genes implicated as eQTL candidates.

4.2 Methodology overview

4.2.1 Selection of candidate reference genes from microarray data

A subset of normal mucosa samples (n=44) collected in the early part of this project was analysed on the Illumina HT12 gene expression microarray as described in 2.3. Quantile-normalised and log-transformed data from this sample set was examined with reference to the genes conventionally used as endogenous controls in human gene expression studies. The maximum fold change (MFC) and the coefficient of variation (CV) for each of these genes were then calculated as measures of expression stability.

4.2.2 qRT-PCR validation on representative samples

Based on the microarray gene expression data, the ten most stable candidate reference genes were shortlisted for a pilot study on representative samples. For each candidate reference gene, gene expression TaqMan® probe and primers were purchased from Life Technologies (Table 4.1). qRT-PCR was performed as described in 2.3.5. Linearity and amplification efficiency of each of these assays were first tested using serial dilutions of HCT116 cell line cDNA.

Gene symbol	Assay ID	Context sequence
PUM1	Hs00472881_m1	TGGGGAACATCAGATCATTCAGTTT
ACTB	Hs99999903_m1	CCTTTGCCGATCCGCCGCCCGTCCA
RPL37A	Hs01102345_m1	GGTGCCTGGACGTACAATACCACTT
PGK1	Hs00943178_g1	AGCCCACAGCTCCATGGTAGGAGTC
UBC	Hs00824723_m1	GTGATCGTCACTTGACAATGCAGAT
ABL1	Hs01104728_m1	GCGAGCATGTTGGCAGTGGAATCCC
EIF2B1	Hs00426752_m1	ATCAAAGATGGAGCGACAATATTGA
RPS17	Hs00734303_g1	GCTGAAGCTTTTGGACTTCGGCAGT
TBP	Hs00427620_m1	GCAGCTGCAAAATATTGTATCCACA
RPL30	Hs00265497_m1	TATCATTGATCCAGGTGACTCTGAC

Table 4.1. TaqMan® assay IDs of candidate reference genes.

Next, the transcript abundance of the 10 candidate genes were measured in representative samples that comprised of cDNA from normal mucosa (n=11), tumour (n=9), ex-vivo normal mucosa (n=8), and colorectal cancer cell lines (n=20). Normal mucosa and tumour tissue were paired whenever possible, and samples were balanced by gender and anatomical site. Ex-vivo normal mucosa samples consisted of untreated samples and samples treated with BMP4 or TGF-β in culture. CRC cell lines consisted of 7 commonly used cell lines (SW480, HCT116, DLD1, HT29, VACO425, COLO320, CACO2) that were untreated or treated with BMP4, TGF-β, aspirin, lithium chloride or retinoic acid. These treatments were performed as I had initially proposed to study the collective effect of CRC risk variants on the TGF-β signalling pathway, however, the study evolved to focus on individual variants as the preliminary results did not demonstrate a collective effect on TGF-β target genes.

4.2.3 Stability ranking of candidate reference genes

The raw CT values from the qRT-PCR were analysed using RefFinder (URL4.1), an online software tool that integrates four major computational programs (BestKeeper [Pffafl *et al*, 2004], GeNorm [Vandesompele *et al*, 2002], Normfinder [Anderson *et al*, 2004], and the comparative delta-Ct method [Silver *et al*, 2006]) to compare and rank the stability of candidate reference genes. Based on the rankings from each program, an appropriate weight is assigned to each individual gene and the geometric mean of their weights is calculated for the overall final ranking.

4.3 Results

4.3.1 Candidate reference genes for pilot qRT-PCR study

There are 32 genes conventionally used as reference genes for human gene expression studies, with commercially available assays (ABI) for each of them (URL4.2 and URL4.3). Of these, 30 genes had matched annotated probes on the Illumina HT12 microarray platform (Table 4.2). A candidate reference gene was defined as a gene with an MFC of less than 2 and a small CV (de Jonge *et al*, 2007). 10 genes had a maximum fold change of >2 (Figure 4.1), hence were considered unsuitable and excluded from further ranking. The remaining 20 genes were then ranked according to their CV, with the 10 most stable genes selected for the pilot study (Table 4.3).

Gene	Description	Annotated probes on HT12
18S	18S ribosomal RNA	No
ABL1	c-abl oncogene 1, non-receptor tyrosine kinase	Yes
ACTB	actin, beta	Yes
B2M	beta-2-microglobulin	Yes
CASC3	cancer susceptibility candidate 3	Yes
CDKN1A	cyclin-dependent kinase inhibitor 1A (p21, Cip1)	Yes
CDKN1B	cyclin-dependent kinase inhibitor 1B (p27, Kip1)	Yes
EIF2B1	eukaryotic translation initiation factor 2B, subunit 1 alpha, 26kDa	Yes
ELF1	E74-like factor 1 (ets domain transcription factor)	Yes
GADD45A	growth arrest and DNA-damage-inducible, alpha	Yes
GAPDH	glyceraldehyde-3-phosphate dehydrogenase	Yes
GUSB	glucuronidase, beta	Yes
HMBS	hydroxymethylbilane synthase	Yes
HPRT1	hypoxanthine phosphoribosyltransferase 1	Yes
IPO8	importin 8	Yes
MRPL19	mitochondrial ribosomal protein L19	Yes
MT-ATP6	mitochondrially encoded ATP synthase 6	No
PES1	pescadillo homolog 1, containing BRCT domain (zebrafish)	Yes
PGK1	phosphoglycerate kinase 1	Yes
POLR2A	polymerase (RNA) II (DNA directed)	Yes
POP4	processing of precursor 4, ribonuclease P/MRP subunit (S. cerevisiae)	Yes
PPIA	peptidylprolyl isomerase A (cyclophilin A)	Yes
PSMC4	proteasome (prosome, macropain) 26S subunit, ATPase, 4	Yes
PUM1	pumilio homolog 1 (Drosophila)	Yes
RPL30	ribosomal protein L30	Yes
RPL37A	ribosomal protein L37a	Yes
RPLP0	ribosomal protein, large, P0	Yes
RPS17	ribosomal protein S17	Yes
TBP	TATA box binding protein	Yes
TFRC	transferrin receptor (p90, CD71)	Yes
UBC	ubiquitin C	Yes
YWHAZ	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, zeta polypeptide	Yes

Table 4.2. 32 conventional human reference genes with commercial assays available.

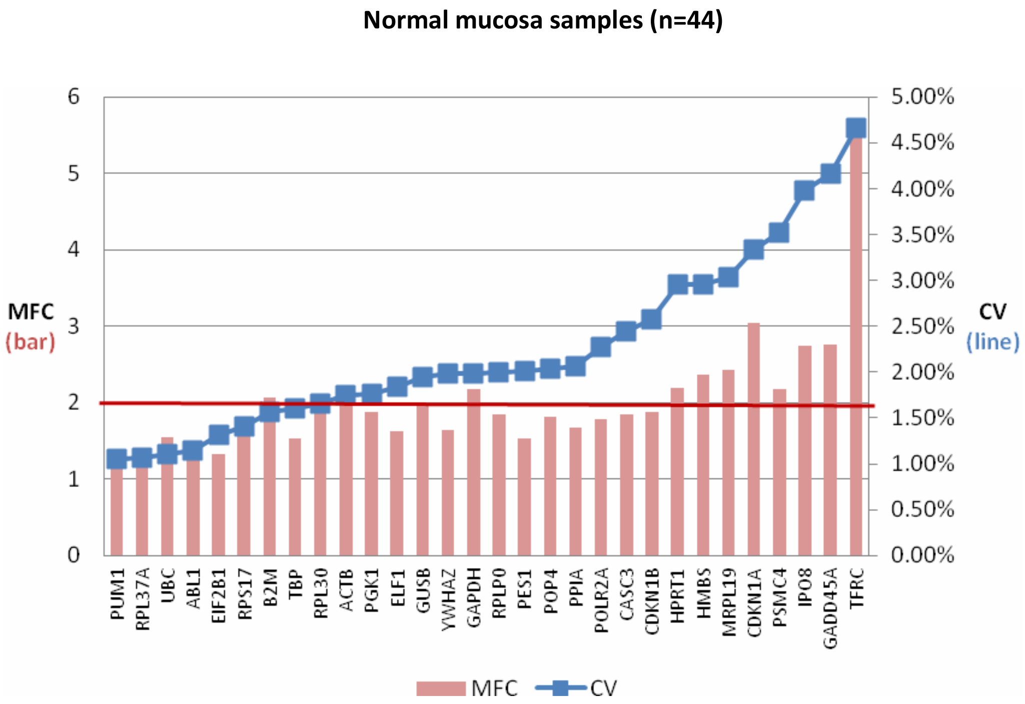


Figure 4.1. Maximum fold change (MFC) and coefficient of variation (CV) of the expression of 30 conventional reference genes in normal mucosa samples. Genes with an MFC>2 (red line represents MFC cut-off) are considered unsuitable as reference genes.

Rank	Gene symbol	MeanExp	SD	CV	MFC
1	***PUM1***	9.589	0.102	1.06%	1.36
2	***RPL37A***	10.128	0.108	1.07%	1.38
3	***UBC***	13.522	0.149	1.10%	1.55
4	***ABL1***	7.320	0.084	1.15%	1.26
5	***EIF2B1***	8.195	0.108	1.32%	1.32
6	***RPS17***	12.334	0.174	1.41%	1.79
Excluded	***B2M***	13.413	0.209	1.56%	2.07
7	***TBP***	7.865	0.126	1.60%	1.53
8	***RPL30***	12.699	0.211	1.66%	1.97
9	***ACTB***	12.884	0.226	1.75%	1.97
10	***PGK1***	9.793	0.173	1.76%	1.88
11	***ELF1***	9.506	0.175	1.84%	1.63
12	***GUSB***	8.970	0.175	1.95%	1.96
13	***YWHAZ***	9.519	0.189	1.98%	1.64
Excluded	***GAPDH***	11.404	0.226	1.98%	2.18
14	***RPLP0***	10.852	0.217	2.00%	1.84
15	***PES1***	7.194	0.145	2.01%	1.53
16	***POP4***	7.758	0.159	2.04%	1.81
17	***PPIA***	7.440	0.153	2.06%	1.68
18	***POLR2A***	9.423	0.215	2.28%	1.79
19	***CASC3***	7.990	0.195	2.44%	1.85
20	***CDKN1B***	8.611	0.222	2.58%	1.88
Excluded	***HPRT1***	8.843	0.261	2.95%	2.20
Excluded	***HMBS***	7.787	0.230	2.95%	2.37
Excluded	***MRPL19***	8.620	0.262	3.03%	2.43
Excluded	***CDKN1A***	11.021	0.367	3.33%	3.05
Excluded	***PSMC4***	7.564	0.266	3.52%	2.18
Excluded	***IPO8***	7.856	0.312	3.98%	2.75
Excluded	***GADD45A***	8.119	0.338	4.16%	2.76
Excluded	***TFRC***	11.222	0.523	4.66%	5.64

Table 4.3. Conventional reference genes ranked according to their CV of expression in 44 normal mucosa samples. Genes with an MFC >2 were considered unsuitable and excluded from stability ranking. (MeanExp=mean of expression, SD=standard deviation, CV=coefficient of variation, MFC=maximum fold change)

4.3.2 qRT-PCR pilot study of candidate reference genes

The mean raw Ct values of the 10 candidate reference genes for each of the sample subgroup are quantified and calculated (Table 4.4). There is a large range of expression levels across these 10 genes, with RPS17 being most highly expressed and TBP being the least expressed.

	ABL1	***ACTB***	***EIF2B1***	***RPL30***	***PGK1***	***PUM1***	***RPS17***	***TBP***	***RPL37A***	***UBC***
NM	27.15	23.73	29.44	26.17	25.42	28.25	22.51	30.70	24.18	24.56
T	26.68	22.80	28.56	25.43	24.12	27.54	21.81	29.80	23.08	24.07
ENM	26.95	24.26	30.10	25.47	24.77	28.33	23.34	31.32	24.42	23.12
CL	26.87	23.34	28.26	25.93	23.83	27.46	21.57	29.57	22.61	23.75

Table 4.4. Comparison of mean cycle threshold (Ct) value across different sample groups. NM = normal mucosa tissue (n=11); T = colon tumour (n=9); ENM = ex-vivo normal mucosa, untreated or treated with BMP4 or TGFβ in culture (n=8); CL = colorectal cancer cell lines, untreated or treated with BMP4, TGFβ, aspirin, lithium chloride or retinoic acid (n=20).

The stability of these candidate reference genes are then analysed and ranked for each tissue subgroup individually, as well as all together (Table 4.5).

STABILITY RANKING	**NM** **(N=11)**	**TUMOUR** **(N=9)**	**EX-VIVO NM** **(N=8)**	**CRC CELL LINES(N=20)**	**ALL** **(N=48)**
1	*EIF2B1*	*TBP*	*PUM1*	*EIF2B1*	*ABL1*
2	*TBP*	*RPL37A*	*TBP*	*RPL30*	*EIF2B1*
3	*UBC*	*EIF2B1*	*ABL1*	*ABL1*	*ACTB*
4	*RPL30*	*UBC*	*EIF2B1*	*ACTB*	*RPL30*
5	*PGK1*	*PUM1*	*RPL30*	*RPL37A*	*RPL37A*
6	*ABL1*	*ACTB*	*RPS17*	*PUM1*	*PUM1*
7	*PUM1*	*RPL30*	*RPL37A*	*PGK1*	*PGK1*
8	*RPS17*	*ABL1*	*PGK1*	*TBP*	*TBP*
9	*ACTB*	*PGK1*	*UBC*	*RPS17*	*RPS17*
10	*RPL37A*	*RPS17*	*ACTB*	*UBC*	*UBC*

Table 4.5. Stability ranking of the candidate endogenous genes in a representative sample set. NM=normal mucosa.

Guided by these results, *EIF2B1*, *TBP* and *RPL30* were selected as reference genes for future qRT-PCR experiments on normal mucosa samples. Although *UBC* was ranked 3rd in the normal mucosa samples, it was ranked 9th in the ex-vivo normal mucosa samples, and was the least stably expressed gene in CRC cell lines. As many of the samples in these 2 groups have been treated with variable courses of BMP4/TGFβ, it is possible that perturbations of the SMAD signalling pathway have an effect on *UBC* transcript levels. Consequently, *UBC* was not selected as a control to avoid any bias during the normalisation of target genes in normal mucosa, many of which may be involved in, or targeted by, SMAD signalling.

EIF2B1, *TBP* and *RPL30* are also well-ranked in the other subgroups and when all samples were analysed altogether; at least two of the three genes were consistently in the top 5 most stable genes across all subgroups. This makes them good reference genes for experiments requiring intra- and inter- subgroup comparison of differential expression.

4.4 Discussion

Due to the speed, sensitivity and specificity it offers, qRT-PCR is widely used in molecular diagnostics, life sciences, agriculture and medicine to quantify gene expression (Bustin *et al*, 2000; Kubista *et al*, 2006). It is one of the key experimental methods used in my investigation of gene expression in the context of common genetic variation associated with colorectal cancer. As these common variants have small effect sizes on cancer risk, it is thought that the associated differential expression, if any, is likely to also be small. Hence, rigorous selection of stably expressed reference genes for normalisation is essential to remove experimentally-induced errors and optimise the detection of subtle variation.

Examining expression microarray data from the tissue type of interest to identify putative reference genes has become a popular approach in our era of high-throughput cell biology (Popovici *et al*, 2009; de Jonge *et al*, 2007). By using this approach, I selected the ten most stable candidates from thirty-two commonly used reference genes for further qRT-PCR validation. These ten putative reference genes represent different cellular processes; the ribosomal proteins RPL30, RPL37A and RPS17, and EIF2B1 are involved in protein synthesis, TBP is a general transcription factor, ACTB is a structural cytoskeletal protein, PGK1 is a glycolytic enzyme, PUM1 is an RNA-binding protein, UBC is involved in ubiquitination and ABL1 is a tyrosine kinase with a role in many key processes linked to cellular growth and survival. Using a set of genes representing a variety of cellular functions is ideal for a pilot experiment, as this means any perturbations resulting from experimental conditions are less likely to affect all the putative reference genes. It is interesting to point out that *GAPDH*, a commonly used reference gene, is relatively unstable within this data set with a MFC of 2.18. This is consistent with previously published reports where variations in *GAPDH* expression have been observed in qRT-PCR experiments (Barber *et al*, 2005; Harper *et al*, 2003). This reinforces the notion that the expression of control genes can vary depending on tissue-type and experimental conditions, hence careful consideration is required before using them for normalisation.

Once putative reference genes have been identified and qRT-PCR validation performed, the circular problem of evaluating the stability of these genes can be resolved using the aforementioned mathematical algorithms. As there is considerable variation in the correlation among the various algorithms (Jacob *et al*, 2013), I utilised a web-based tool that integrates the four most commonly used approaches to assess the stability of these genes. It quickly becomes apparent that the expression variability of these genes differs between the tissue subgroups. Though unsurprising, as heterogeneity is more likely in the tumour and cell line subgroups as compared to normal mucosa, this again demonstrates that expression variation of reference genes is present and dependent, at least in part, on the tissue type. More importantly, the difference in the normal mucosa and ex-vivo normal mucosa subgroup suggests that perturbations of the SMAD signalling pathway could systematically affect expression of these 'housekeeping' genes. Given that the SMAD signalling pathway is frequently altered in cancer, it is all the more crucial that reference genes are validated when the experimental setup involves comparison between normal and tumour tissue. A good post-hoc experiment would be to compare the expression of a gene of interest normalised with the most stable reference gene, and the same gene normalised with the least stable reference gene. Any demonstrable difference would consolidate the importance of rationalising reference genes prior to their use for normalisation.

It is strongly recommended that three or more reference genes are used to reduce the effect of any variation in a single reference gene. However, it is recognised that this may not always be cost-effective, especially in experiments where the expected fold change in the expression of the gene of interest is substantially larger than any potential variation in the reference gene expression. In these instances, the need for normalisation accuracy may have to be weighed up against the practical constraints of time and resources. Such is the case with my siRNA knockdown studies in cell lines, where there is 75-90% knockdown of the target genes (Figure 8.13). The use of a single reference gene may not be ideal but on this occasion we have accepted it as a caveat, given that the differential expression of the genes of interest is appreciably larger than the fluctuation of the reference gene expression. Other biological factors may also influence the choice of the reference genes, and should be carefully considered when the information is available. For example, one of my experiments involved quantifying the knockdown of NF-Y subunits, which is a ubiquitous transcription-factor that is known to bind to variety of genes (7.3.4). In this instance, TBP would be a good reference gene as it is known to lack the CCAAT consensus sequence for NF-Y binding (Nardini *et al*, 2013), and will be more likely than other reference genes to be unaffected by the depletion of NF-Y. Hence, a combination of biological justification and validation with a pilot experiment provides the ideal platform for the informed selection of qRT-PCR reference genes.

In conclusion, the reference genes *EIF2B1*, *TBP* and *RPL30* were chosen to validate any differential expression as identified by microarray analysis of the colorectal normal mucosa samples.

Chapter 5

Gender- and site-specific differential gene expression in the human colorectal normal mucosa

5.1 Introduction

To systematically characterise the functional effect of low-penetrance common genetic variation that are associated with susceptibility to colorectal cancer, whole-genome gene expression microarray analysis was undertaken to look for evidence of association with gene expression. Before analysing the gene expression data in relation to genetic risk variants, several clinical variables were first examined in relation to global gene expression, namely age, gender, cancer status and anatomical sampling-site. Although the relevance of this to the functional characterisation of risk variants may not be immediately apparent, it would provide insight into the regulation and heterogeneity of gene expression in the normal colorectal mucosa. Positive findings would inform the design of the risk variant analyses to minimise confounding effects of these variables, hence optimising the detection of subtle effects that are associated with inherited variation. Examining these variables that are known to affect gene expression levels will also provide a form of internal validation of the microarray platform used in this study.

Age- and gender- related differences in gene expression in the colon is relevant for two reasons. Firstly, colorectal cancer incidence is higher in men than women and strongly increases with age (as reviewed by Brenner *et al*, 2014). Secondly, previous studies have indicated that there are age-related changes in gene expression levels occurring in various human tissue types (Somel *et al*, 2006; Glass *et al*, 2013), as well as sexual dimorphism in non-reproductive tissues (reviewed by Rinn *et al*, 2006). Although this has not been specifically demonstrated in the human large intestine, there is evidence that expression of some genes are gender-biased in the oesophageal (Menon *et al*, 2011) and small intestinal mucosa (Sankaran-Walters *et al*, 2013).

The concept of field cancerisation has been proposed to explain the development of multiple primary tumours in the same organ and locally recurrent cancer in patients who have multifocal cancer without apparent familial predisposition. Providing evidence for this is a study describing methylation of the *MGMT* gene promoter in normal-appearing colorectal mucosa adjacent to colorectal cancer with *MGMT* promoter methylation (Shen *et al*, 2005). The proximity to the tumour appeared to influence the likelihood of hypermethylation in the normal mucosa. In view of this, the normal mucosa samples in my study were collected from the resection margin furthest from the tumour to minimise any field effect that may be present, and the presenting pathology of the donor patients were categorised as cancer or non-cancer for gene expression analysis.

There are known epidemiological, clinical and molecular differences between proximal and distal colon tumours, suggesting that the risk factors and transformation pathways of sporadic colorectal cancer differ according to the anatomical location within the colon (as reviewed by Iacopetta *et al*, 2002). This may reflect the different biological characteristics of the normal colorectal mucosa within the different segments, or a segmental heterogeneity in the gut environment i.e. the microbiome and metabolites, or most likely, the interaction between a distinct environment and distinct target cells. The known site-specific differences in normal and cancerous conditions of the colon is summarised in Table 5.1. In view of this, the sampling site for each of the normal mucosa samples collected for this study were classified as proximal or distal to the splenic flexure (Figure 5.1) and examined for differential expression.

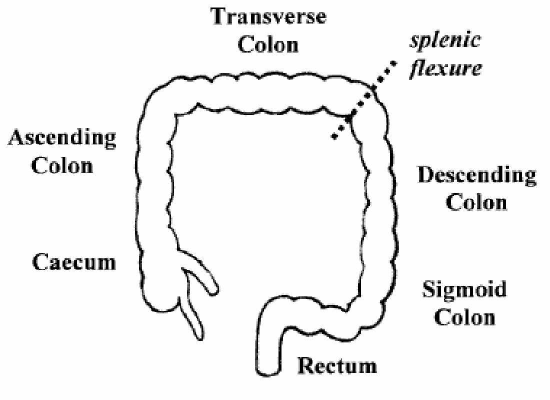


Figure 5.1 Schematic drawing of the human colon illustrating the splenic flexure cut-off point, which determines the proximal/distal classification used in this study. (Figure adapted from Iacopetta, 2002)

A)

NORMAL	***Proximal colon***	***Distal colon***
Development	Embryonic midgut	Embryonic hindgut
Vascular supply	Superior mesenteric artery	Inferior mesenteric artery
Average crypt length	Shorter	Longer
Short-chain fatty acid production by fermentation	8-fold higher	
Mutagenic metabolites		Higher exposure
Metabolism of bile acids	Higher	
Activity of ornithin decarboxylase		Higher
Methylation of ER gene	Higher	

B)

CANCER	***Proximal colon***	***Distal colon***
Mismatch repair defective (MSI-H)	30%	2%
Mucinous tumours	Frequent	Infrequent
Familial cancer syndromes	Lynch Syndrome	FAP
Karyotype	Pseudo-diploid	Aneuploid with LOH
Gene mutations (TP53 and K-RAS) and C-MYC expression	Lower frequency	Higher frequency
5-FU chemotherapy response	Good	Marginal or none
Gender	Proportion of cancer in the distal colon is lower among women than among men	

Table 5.1 Summary of the differences between the proximal and distal colon, in both normal conditions and neoplastic disease. (Adapted from Iacopetta *et al*, 2002; Glebov *et al*, 2003)

5.2 Methodological overview

5.2.1 Study subjects and biological material

115 normal colorectal mucosa samples were collected from patients undergoing large bowel resections as described in 2.1. The characteristics of the study subjects are summarised in Table 5.2 and detailed in Table 5.3. Study subjects were categorised by age (≤60 or >60), gender, presenting pathology (cancer vs without cancer), and the anatomical sampling site (proximal vs distal).

Age	**≤60**	**>60**
	28	87
Gender	**Males**	**Females**
	64	51
Presenting pathology	**Cancer**	**Without cancer**
	99	16
Anatomical sampling site	**Proximal colon**	**Distal colon**
	39	76

Table 5.2 Summary of the characteristics of patients who donated tissue samples for this study.

Study ID	***Gender***	***Age***	***Site***	***Condition***	***Matched PBMC***
2335	M	61	Proximal	Adenocarcinoma	Yes
11481	M	82	Distal	Adenocarcinoma	Yes
11559	F	71	Distal	Adenocarcinoma	Yes
11588	F	88	Distal	Adenocarcinoma	No
11868	F	77	Proximal	Adenocarcinmoa	No
11913	M	51	Distal	Adenocarcinoma	Yes
11915	F	86	Proximal	Adenocarcinoma	No
11981	F	77	Distal	Diverticular disease	No
11986	F	73	Distal	Tubulo-villous adenoma	No
11990	M	63	Distal	Diverticulitis	No
12002	M	77	Proximal	Adenocarcinoma	No
12003	F	65	Distal	Adenocarcinoma	Yes
12032	F	75	Proximal	Adenocarcinoma	No
12033	M	73	Proximal	Adenocarcinoma	Yes
12037	M	80	Distal	Adenocarcinoma	Yes
12039	F	57	Distal	Adenocarcinoma	No
12040	M	56	Proximal	Adenocarcinoma	Yes
12041	F	67	Proximal	Adenocarcinoma	Yes
12042	F	75	Proximal	Adenocarcinoma	Yes
12046	F	57	Distal	Adenocarcinoma	No
12047	F	79	Proximal	Adenocarcinoma	No
12048	F	64	Distal	Adenocarcinoma	Yes
12049	M	69	Distal	Adenocarcinoma	Yes
12050	F	74	Distal	Adenocarcinoma	No
12051	M	71	Proximal	Tubulo-villous adenoma	No
12052	M	61	Distal	Tubulo-villous adenoma	No
12053	F	79	Proximal	Adenocarcinoma	No
12054	M	66	Distal	Villous adenoma	No
12056	F	71	Proximal	Adenocarcinoma	No
12057	M	68	Distal	Adenocarcinoma	No
12059	M	70	Distal	Adenocarcinoma	Yes
12061	M	62	Distal	Adenocarcinoma	No
12063	M	58	Distal	Adenocarcinoma	No
12064	F	67	Distal	Adenocarcinoma	Yes
12065	F	79	Distal	Adenocarcinoma	Yes
12067	M	62	Distal	Adenocarcinoma	Yes
12068	F	75	Proximal	Tubulo-villous adenoma	No
12069	F	80	Proximal	Adenocarcinoma	No
12070	F	71	Proximal	Adenocarcinoma	No
12071	M	67	Distal	Adenocarcinoma	Yes

12100	M	58	Distal	Adenocarcinoma	No
12114	M	64	Proximal	Adenocarcinoma	No
12147	F	47	Distal	Adenocarcinoma	Yes
12202	M	42	Distal	Adenocarcinoma	Yes
12208	F	87	Proximal	Adenocarcinoma	Yes
12234	M	47	Distal	Adenocarcinoma	Yes
12236	M	76	Distal	Adenocarcinoma	Yes
12253	M	65	Proximal	Adenocarcinoma	Yes
12254	M	67	Proximal	Adenocarcinoma	No
12255	F	74	Proximal	Adenocarcinoma	Yes
12256	F	57	Distal	Adenocarcinoma	Yes
12259	M	75	Proximal	Adenocarcinoma	Yes
12260	F	57	Distal	Adenocarcinoma	No
12272	M	65	Distal	Adenocarcinoma	No
12304	M	59	Proximal	Adenocarcinoma	Yes
12305	M	63	Distal	Adenocarcinoma	Yes
12307	F	67	Distal	Adenocarcinoma	No
12312	F	42	Distal	Intestinal dysmobility	No
12316	M	77	Proximal	Villous adenoma	No
12369	M	80	Distal	Adenocarcinoma	Yes
12370	M	41	Distal	Diverticular disease	No
12407	F	64	Distal	Adenocarcinoma	Yes
12408	M	67	Distal	Adenocarcinoma	Yes
12409	M	50	Distal	Adenocarcinoma	No
12412	F	75	Proximal	Adenocarcinoma	Yes
12415	M	80	Distal	Adenocarcinoma	No
12419	M	65	Proximal	Adenocarcinoma	Yes
12421	F	59	Distal	Adenocarcinoma	No
12433	M	59	Distal	Adenocarcinoma	Yes
12435	M	83	Distal	Adenocarcinoma	Yes
12451	M	79	Distal	Tubulo-villous adenoma	No
12454	F	64	Distal	Diverticular disease	No
12464	F	72	Proximal	Adenocarcinoma	No
12468	M	55	Distal	Adenocarcinoma	Yes
12473	M	86	Distal	Adenocarcinoma	Yes
12475	M	80	Distal	Adenocarcinoma	Yes
12477	F	65	Distal	Adenocarcinoma	Yes
12481	F	39	Distal	Adenocarcinoma	Yes
12483	F	54	Distal	Adenocarcinoma	Yes
12519	M	47	Distal	Adenocarcinoma	Yes
12520	F	60	Proximal	Adenocarcinoma	Yes
12529	M	62	Distal	Adenocarcinoma	Yes

12555	M	69	Distal	Tubulo-villus adenoma	No
12562	F	52	Distal	Diverticular disease	No
12568	M	69	Proximal	Tubulo-villous adenoma	No
12584	M	48	Distal	Adenocarcinoma	No
12586	M	62	Distal	Adenocarcinoma	Yes
12587	M	63	Proximal	Adenocarcinoma	No
12597	M	84	Proximal	Adenocarcinoma	Yes
12602	M	73	Distal	Adenocarcinoma	Yes
12609	M	71	Proximal	Adenocarcinoma	Yes
12619	F	80	Distal	Adenocarcinoma	Yes
12624	F	33	Distal	Adenocarcinoma	Yes
12630	M	69	Proximal	Adenocarcinoma	Yes
12631	M	69	Distal	Adenocarcinoma	Yes
12633	M	75	Distal	Adenocarcinoma	No
12634	F	66	Proximal	Adenocarcinoma	Yes
12645	F	27	Distal	Adenocarcinoma	No
12646	F	69	Distal	Adenocarcinoma	Yes
12647	M	82	Proximal	Adenocarcinoma	Yes
12650	M	65	Proximal	Adenocarcinoma	Yes
12659	M	71	Distal	Adenocarcinoma	Yes
12660	M	68	Distal	Adenocarcinoma	Yes
12668	M	79	Distal	Adenocarcinoma	No
12669	F	75	Distal	Adenocarcinoma	Yes
12726	F	60	Distal	Adenocarcinoma	No
12741	F	71	Proximal	Adenocarcinoma	No
12751	F	79	Distal	Adenocarcinoma	No
12775	F	61	Distal	Adenocarcinoma	No
12779	M	73	Distal	Adenocarcinoma	No
12810	M	63	Distal	Adenocarcinoma	No
12812	M	52	Distal	Adenocarcinoma	No
12813	M	78	Distal	Adenocarcinoma	No
12854	F	67	Proximal	Adenocarcinoma	No
12856	F	74	Proximal	Tubulo-villous adenoma	No

Table 5.3 Characteristics of the 115 study subjects – age, gender, sampling site (proximal or distal colon), the indication for bowel resection surgery and whether matched PBMC were collected.

5.2.2 Gene expression profiling and analysis

RNA was extracted, amplified and hybridised on Illumina HT12 gene expression microarrays as described in 2.3. The expression profiles of the 115 normal mucosa samples were analysed using the R *limma* package as described in 2.6.2, using the co-variates age, gender, presenting pathology and anatomical sampling site as categorised in Table 5.2.[3]

5.2.3 qRT-PCR validation

Technical validation was performed with qRT-PCR as described in 2.3. *EIF2B1*, *RPL30* and *TBP* were selected as reference genes for normalisation as described in Chapter 4. The Taqman® Gene Expression assays for the genes of interest are listed in Table 5.4.

Gene symbol	Assay ID	Context sequence
PRAC	Hs00741541_g1	AGAGTGCTTTTCTCTCTAATAAGAA
PITX2	Hs04234069_mH	GAGTCCGGGTTTGGTTCAAGAATCG
CKB	Hs01058288_g1	CCTCACCCAGATTGAAACTCTCTTC
OLFM4	Hs00197437_m1	TCCCACTCCAGGGAGCTGTGGTCAT
L1TD1	Hs00219458_m1	TTTTTCGCCAGGCACCAAGGCACAG

Table 5.4 TaqMan® assay IDs of the genes of interest in this chapter.

[3] Analysis performed by Graeme Grimes, MRC Human Genetics Unit, IGMM.

5.3 Results

5.3.1 No detectable differential expression in the normal mucosa between age groups and presenting pathology

Approximately 29,000 probes were detected in the normal mucosa (*p*-value<0.01), mapping to approximately 20,000 unique annotated genes. The expression profiles of the 115 normal mucosa samples were analysed using the R *limma* package and showed no detectable differential expression by age (≤60, n=28 ; >60, n=87) or presenting pathology (with cancer, n=99 ; without cancer, n=16). However, it should be noted that these variables are not well-balanced, reducing the power to detect associated differences.

5.3.2 Gender-specific differential expression of the human colorectal mucosa

23 genes were more highly expressed in males (n=64) whereas 22 genes were more highly expressed in females (n=51). All the genes that were significantly differentially expressed between genders are shown in Tables 5.5 and 5.6.

Gene	FC	AvgExp	FDR *q*-Val	Chr	Description
RPS4Y1	17.55	9.09	8.68E-122	Y	ribosomal protein S4, Y-linked 1
EIF1AY	5.76	8.03	9.70E-100	Y	eukaryotic *trans*lation initiation factor 1A, Y-linked
CYORF15A	2.64	7.42	1.85E-77	Y	taxilin gamma 2, pseudogene on chrY
EIF1AY	1.92	7.09	1.34E-68	Y	eukaryotic *trans*lation initiation factor 1A, Y-linked
JARID1D	2.12	7.16	1.38E-64	Y	lysine (K)-specific demethylase 5D on chrY
LOC643123	1.45	6.79	1.28E-47	Y	arylsulfatase F pseudogene 1
ZFY	1.33	6.65	1.50E-41	Y	zinc finger protein, Y-linked
TMSB4Y	1.51	6.81	7.04E-41	Y	thymosin beta 4, Y-linked
PRKY	1.40	6.83	1.18E-38	Y	protein kinase, Y-linked, pseudogene
RPS4Y2	2.91	7.46	2.39E-37	Y	ribosomal protein S4, Y-linked 2
UTY	1.28	6.74	8.53E-30	Y	ubiquitously *trans*cribed tetratricopeptide repeat containing, Y-linked
USP9Y	1.14	6.54	7.16E-22	Y	ubiquitin specific peptidase 9, Y-linked
TTTY2	1.12	6.58	8.39E-18	Y	testis-specific *trans*cript, Y-linked 2 (non-protein coding)
TTTY14	1.12	6.62	2.40E-13	Y	testis-specific *trans*cript, Y-linked 14 (non-protein coding)
CD99	1.30	8.82	3.53E-10	X	CD99 molecule
DDX3Y	1.09	6.58	3.13E-06	Y	DEAD (Asp-Glu-Ala-Asp) box helicase 3, Y-linked
NLGN4Y	1.05	6.51	1.23E-05	Y	neuroligin 4, Y-linked
OSCP1	1.06	6.68	0.006	1	organic solute carrier partner 1
DPM3	1.11	8.43	0.037	1	dolichyl-phosphate mannosyl*trans*ferase polypeptide 3
MGC72104	1.19	8.30	0.039	Y	FSHD region gene 1 family, member B
CSTF3	1.12	7.93	0.043	11	cleavage stimulation factor, 3' pre-RNA, subunit 3, 77kDa
ZMYND12	1.06	6.81	0.046	1	zinc finger, MYND-type containing 12
LOC729137	1.08	6.67	0.047	Y	zinc finger protein 839-like

Table 5.5 23 unique genes are more highly expressed in males than females, listed in the order of adjusted *p*-value. Where more than one probe is present for a single gene, the probe with the highest *p*-value is presented. FC=fold change as calculated by male/female mean expression, AvgExp=Average expression of gene.

Gene	Fc	AvgExp	FDR *q*-Val	Chr	Description
XIST	4.05	7.74	1.36E-71	X	X inactive specific *trans*cript (non-protein coding)
HDHD1A	1.40	7.99	7.44E-22	X	Haloacid Dehalogenase-Like Hydrolase Domain Containing 1
UTX	1.21	7.38	1.38E-17	X	Ubiquitously *trans*cribed tetratricopeptide repeat, X chromosome
ARSD	1.54	9.03	8.04E-15	X	arylsulfatase D
ZFX	1.11	6.90	1.49E-09	X	zinc finger protein, X-linked
U2AF1L2	1.21	7.81	2.30E-09	X	CCCH Type Zinc Finger, RNA-Binding Motif And Serine/Arginine Rich Protein
TRAPPC2	1.23	8.30	9.70E-09	X	trafficking protein particle complex 2
PNPLA4	1.11	6.74	3.38E-08	X	patatin-like phospholipase domain containing 4
PRKX	1.09	6.94	4.36E-07	X	protein kinase, X-linked
RPS4X	1.30	11.22	7.15E-07	X	ribosomal protein S4, X-linked
ARSE	1.17	7.11	1.64E-06	X	arylsulfatase E (chondrodysplasia 84unctate 1)
ZRSR2	1.07	6.71	1.67E-05	X	zinc finger (CCCH type), RNA-binding motif and serine/arginine rich 2
HEPH	1.23	9.29	3.63E-05	X	hephaestin
DDX3X	1.21	9.80	0.001	X	DEAD (Asp-Glu-Ala-Asp) box helicase 3, X-linked
EIF1AX	1.33	8.32	0.003	X	eukaryotic *trans*lation initiation factor 1A, X-linked
OFD1	1.16	7.87	0.003	X	oral-facial-digital syndrome 1
GYG2	1.12	7.05	0.007	X	glycogenin 2
POF1B	1.11	6.96	0.008	X	premature ovarian failure, 1B
NLRP2	1.17	6.91	0.015	19	NLR family, pyrin domain containing 2
NLGN4X	1.14	7.19	0.019	X	neuroligin 4, X-linked
UBE1	1.17	9.53	0.035	X	ubiquitin-activating enzyme E1
COPS8	1.08	7.26	0.047	2	COP9 signalosome subunit 8

Table 5.6 22 unique genes are more highly expressed in females than males, listed in the order of adjusted *p*-value. Where more than one probe is present for a single gene, the probe with the highest p-value is presented. FC=fold change as calculated by female/male mean expression, AvgExp=Average expression of gene.

5.3.3 Gene expression differences between the human proximal and distal colorectal mucosa

There was differential expression in 1303 probes (1007 unique genes) between the proximal (n=39) and the distal colon (n=76) (*p*-value<0.05, FDR adjusted), with 55 unique genes showing >1.5 fold difference. 29 of these genes were expressed at a higher level in the proximal colon and 26 in the distal colon, as listed in Table 5.7 and Table 5.8.

qRT-PCR validation of these expression differences was performed for five out of twelve genes with a differential fold change of >2, confirming the site-related differential expression observed for *PITX2* (*p*<2.2E-16), *L1TD1* (*p*=5.3E-13), *OLFM4* (*p*=2.7E-07), *PRAC* (*p*<2.2E-16) and *CKB* (*p*=3.6E-07).[4] Highly significant correlations with Spearman Rho values of >0.75 were observed between the two different expression quantification techniques (Figures 5.2-5.6).

[4] qRT-PCR performed under close supervision by undergraduate student Fanny Roth.

Gene	FC	AvgExp	FDR *q*-Val	Description
PITX2	2.83	7.17	7.03E-39	paired-like homeodomain 2
L1TD1	2.01	7.35	4.64E-19	LINE-1 type *trans*posase domain containing 1
ETNK1	2.35	8.46	8.67E-15	ethanolamine kinase 1
SLC23A3	1.66	7.48	1.02E-14	solute carrier family 23, member 3
IGFBP2	1.65	7.91	1.36E-12	insulin-like growth factor binding protein 2, 36kDa
MB	1.55	7.61	3.75E-11	myoglobin
MEP1B	1.56	7.02	1.43E-09	meprin A, beta
SLC20A1	1.82	10.07	3.23E-09	solute carrier family 20 (phosphate *trans*porter), member 1
NQO1	1.57	10.08	3.48E-08	NAD(P)H dehydrogenase, quinone 1
ANPEP	2.17	9.31	6.71E-08	alanyl (membrane) aminopeptidase
ADH1C	1.62	11.13	1.54E-07	alcohol dehydrogenase 1C (class I), gamma polypeptide
OLFM4	3.18	11.15	3.82E-07	olfactomedin 4
OASL	1.62	7.67	1.07E-06	2'-5'-oligoadenylate synthetase-like
FAM3B	1.69	6.73	2.93E-06	family with sequence similarity 3, member B
PROM1	1.62	9.77	1.39E-05	prominin 1
NR1H4	1.56	7.85	3.73E-05	nuclear receptor subfamily 1, group H, member 4
DEFB1	1.57	8.6	8.54E-05	defensin, beta 1
C6ORF105	1.55	9.28	0.000233	androgen-dependent TFPI-regulating protein
OSTALPHA	1.97	8.32	0.000235	Solute carrier family 51, alpha subunit
CCL13	1.74	8.86	0.000345	chemokine (C-C motif) ligand 13
CCL8	1.69	9.23	0.000442	chemokine (C-C motif) ligand 8
UGT2B15	1.75	8.36	0.002264	UDP glucuronosyl*trans*ferase 2 family, polypeptide B15
DEFA5	1.84	7.28	0.006333	defensin, alpha 5, Paneth cell-specific
UGT2B11	1.85	10.08	0.007722	UDP glucuronosyl*trans*ferase 2 family, polypeptide B11
NBPF20	1.52	8.55	0.008969	Neuroblastoma Breakpoint Family Member 20
REG3A	1.56	6.91	0.009994	regenerating islet-derived 3 alpha
UGT2B17	1.84	9.55	0.012681	UDP glucuronosyl*trans*ferase 2 family, polypeptide B17
UGT2B7	1.79	11.7	0.017934	UDP glucuronosyl*trans*ferase 2 family, polypeptide B7
VIP	1.68	9.58	0.021505	vasoactive intestinal peptide

Table 5.7 Genes that are more highly expressed in the proximal colon compared to the distal colon, listed in order of adjusted *p*-value. Where more than one probe is present for a single gene, the probe with the highest *p*-value is presented. FC=fold change as calculated by proximal/distal mean expression, AvgExp=Average expression of gene.

Gene	FC	AvgExp	FDR *q*-Val	Description
PRAC	17.39	9.26	7.76E-46	prostate cancer susceptibility gene 1
ST6GALNAC6	2.19	8.78	6.53E-17	ST6 (alpha-N-acetyl-neuraminyl-2,3-beta-galactosyl-1,3)-N-acetylgalactosaminide alpha-2,6-sialyl*trans*ferase 6
CLDN8	3.29	9.54	1.68E-15	claudin 8
ST3GAL4	1.66	7.36	2.99E-15	ST3 beta-galactoside alpha-2,3-sialyl*trans*ferase 4
HOXB13	1.85	7.13	5.78E-15	homeobox B13
SPON1	1.61	9.38	7.52E-13	spondin 1, extracellular matrix protein
LGALS2	1.85	8.56	1.77E-11	lectin, galactoside-binding, soluble, 2
CAPN13	1.71	8.69	7.30E-11	calpain 13
LOC387882	1.54	8.64	3.92E-09	uncharacterised
CKB	2.06	10.79	5.19E-09	creatine kinase, brain
MUC17	2.10	8.36	1.13E-08	mucin 17, cell surface associated
TFF1	2.51	8.51	4.44E-08	trefoil factor 1
LOC401321	1.65	8.35	6.80E-08	uncharacterised
CRIP1	1.73	10.87	1.58E-07	cysteine-rich protein 1 (intestinal)
WFDC2	1.84	8.58	1.97E-07	WAP four-disulfide core domain 2
KRTAP13-2	1.77	7.12	6.40E-07	keratin associated protein 13-2
SPINK5	1.64	7.50	1.07E-06	serine peptidase inhibitor, Kazal type 5
MUC12	1.93	10.18	1.19E-06	mucin 12, cell surface associated
PYY	1.75	8.63	5.02E-06	peptide YY
TMEM200A	1.52	7.72	2.52E-05	*trans*membrane protein 200A
PI3	2.14	8.65	7.83E-05	peptidase inhibitor 3, skin-derived
GLDN	1.53	6.92	8.54E-05	gliomedin
GCG	1.73	9.07	0.000255	glucagon
SLC28A2	1.57	7.45	0.001927	solute carrier family 28 (concentrative nucleoside *trans*porter), member 2
S100P	1.66	9.40	0.004687	S100 calcium binding protein P
INSL5	1.58	7.12	0.006683	insulin-like 5

Table 5.8 Genes that are more highly expressed in the distal colon compared to the proximal colon, listed in order of adjusted *p*-value. Where more than one probe is present for a single gene, the probe with the highest p-value is presented. FC=fold change as calculated by proximal/distal mean expression, AvgExp=Average expression of gene.

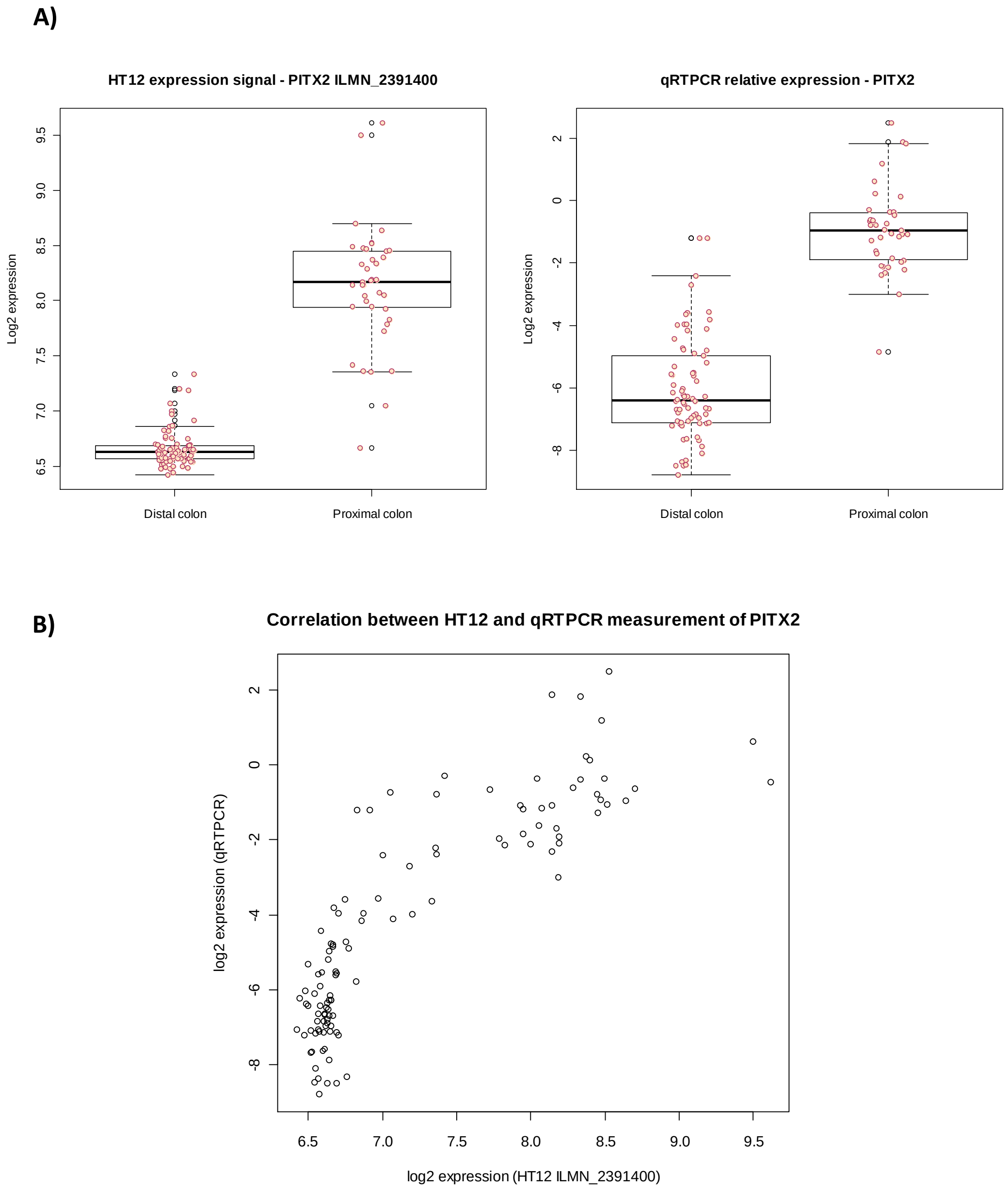


Figure 5.2. **A)** Boxplots of *PITX2* expression as quantified by the Illumina HT12 microarray or qRT-PCR (unpaired t-test, *p*<2.2E-16). **B)** Scatterplot demonstrating the relationship between the expression of *PITX2* as measured by the two different techniques (Spearman rho=0.843, *p*-value=< 2.2e-16).

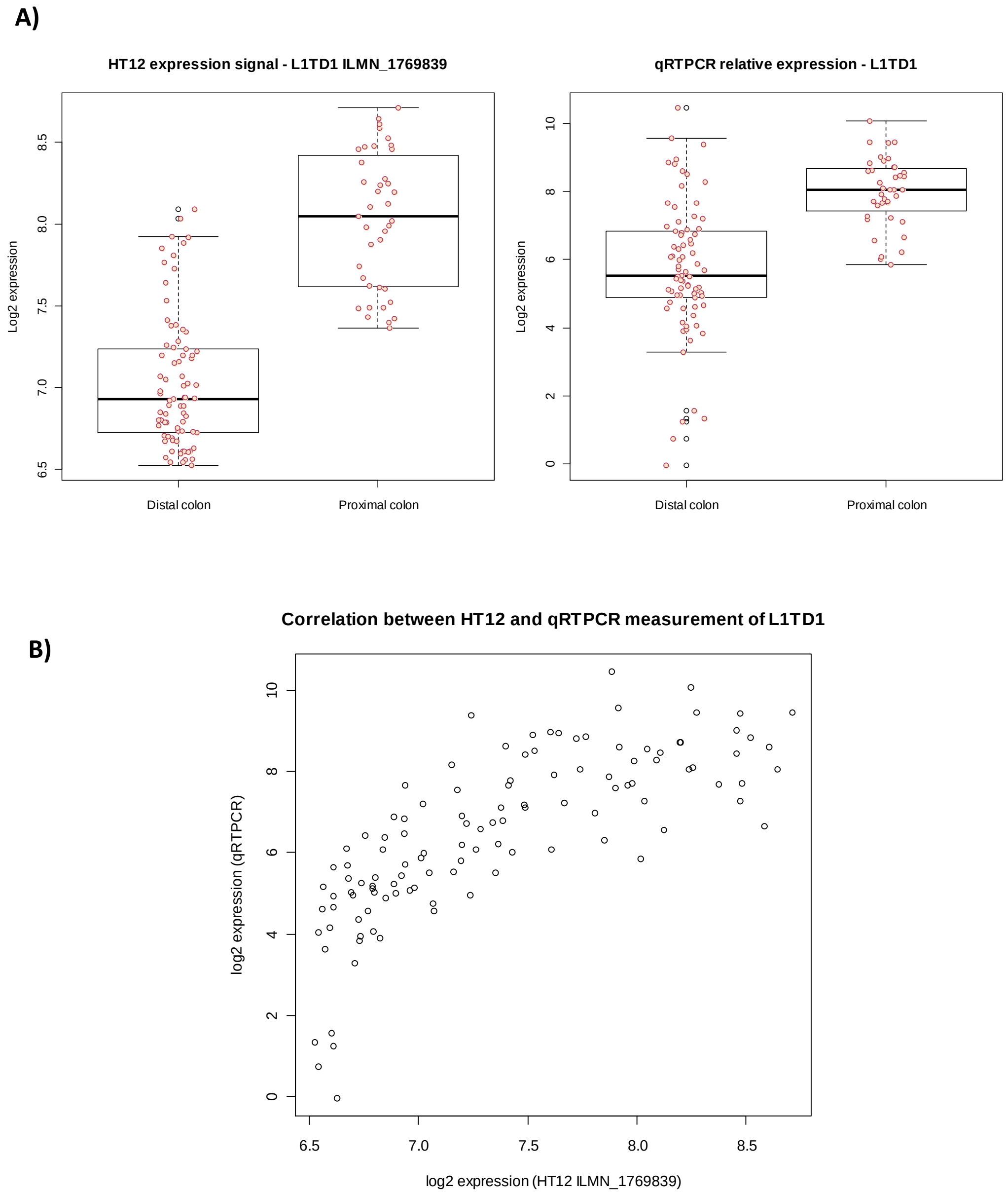


Figure 5.3. **A)** Boxplots of *L1TD1* expression as quantified by the Illumina HT12 microarray or qRT-PCR (unpaired t-test, *p*=5.3E-13). **B)** Scatterplot demonstrating the relationship between the expression of *L1TD1* as measured by the two different techniques (Spearman rho=0.828, *p*-value=< 2.2e-16).

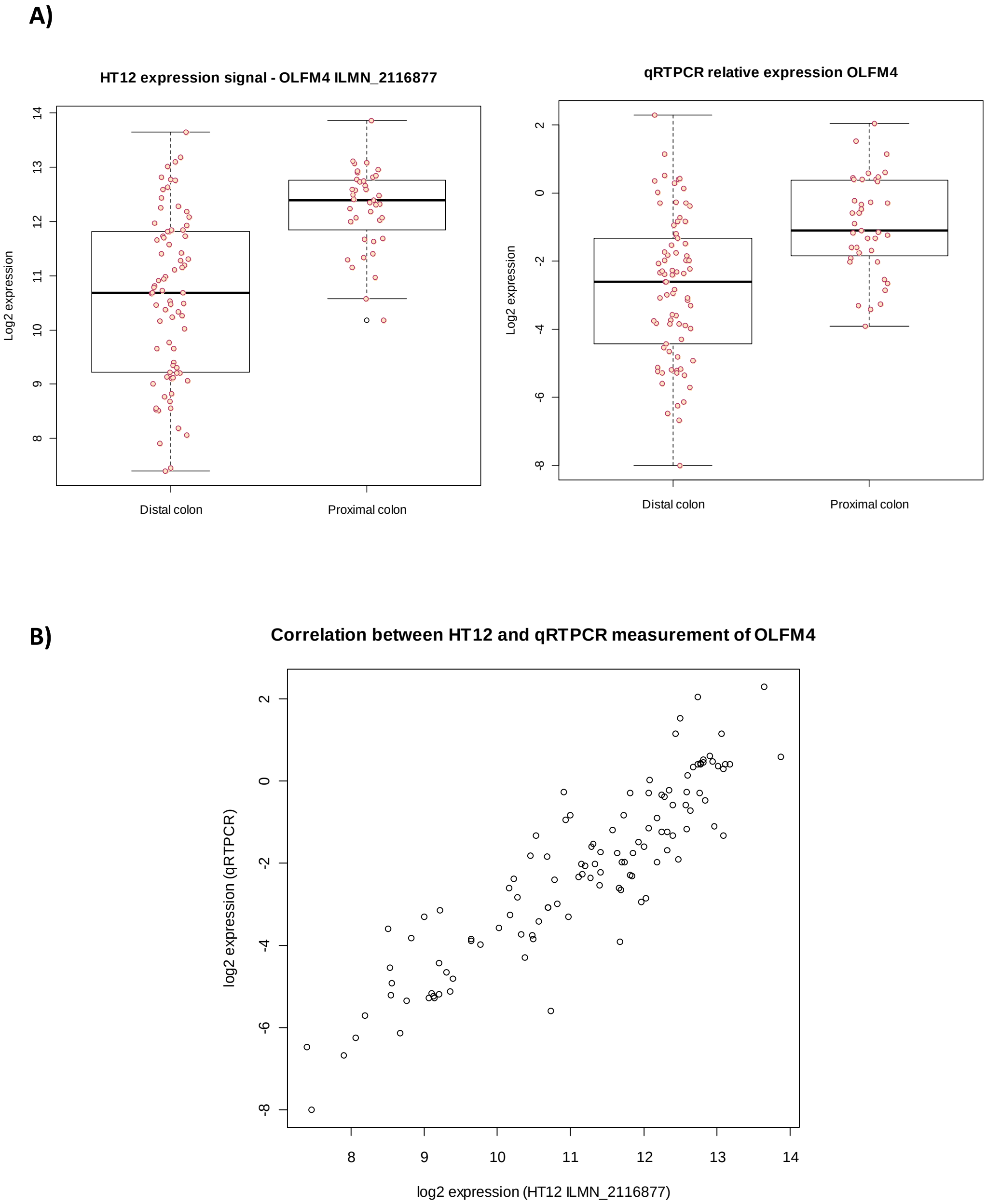


Figure 5.4. **A)** Boxplots of *OLFM4* expression as quantified by the Illumina HT12 microarray or qRT-PCR (unpaired t-test, *p*=2.7E-07). **B)** Scatterplot demonstrating the relationship between the expression of *OLFM4* as measured by the two different techniques (Spearman rho=0.895, *p*-value=< 2.2e-16).

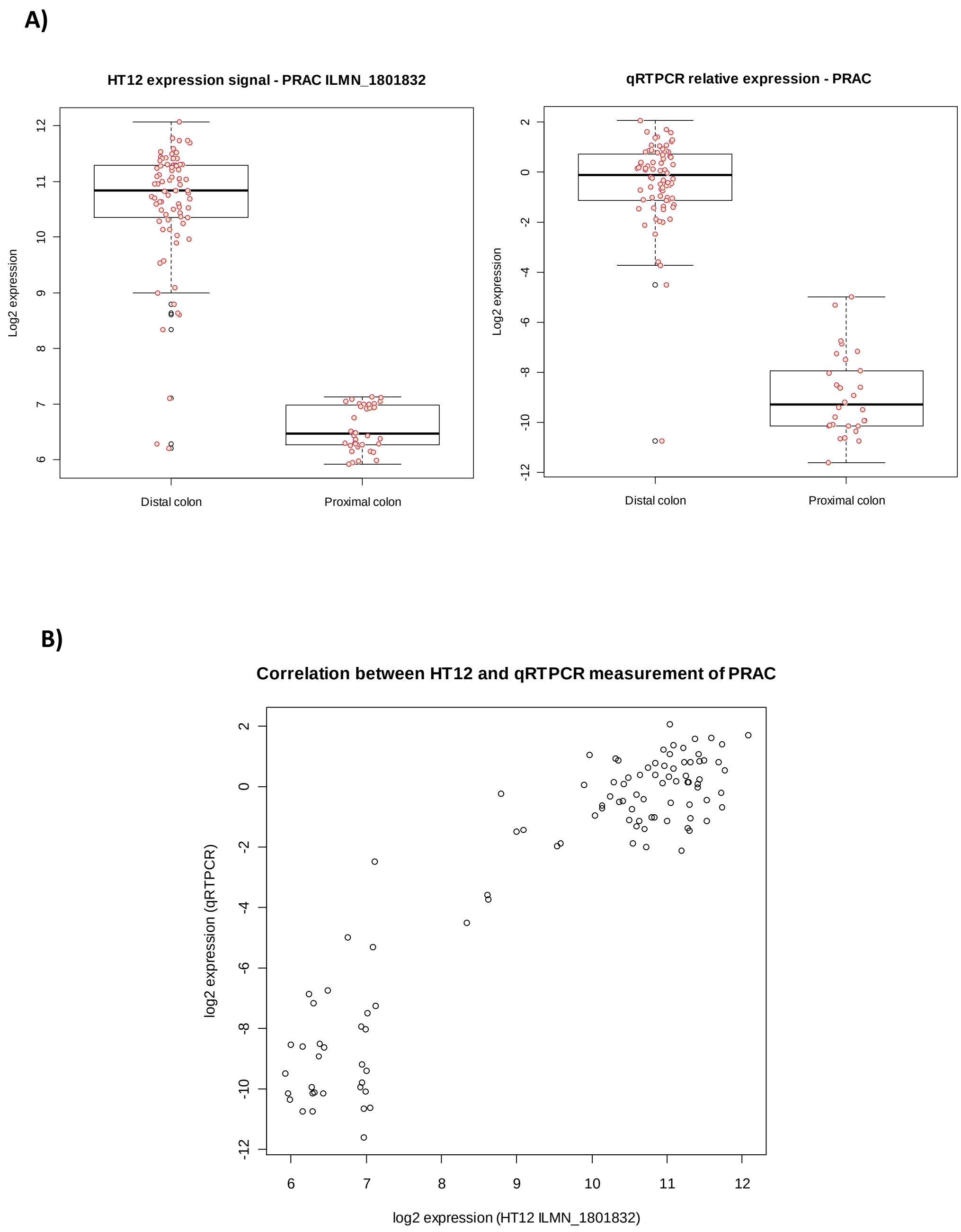


Figure 5.5. **A)** Boxplots of *PRAC* expression as quantified by the Illumina HT12 microarray or qRT-PCR (unpaired t-test, *p*<2.2E-16). **B)** Scatterplot demonstrating the relationship between the expression of *PRAC* as measured by the two different techniques (Spearman rho=0.791, *p*-value=< 2.2e-16).

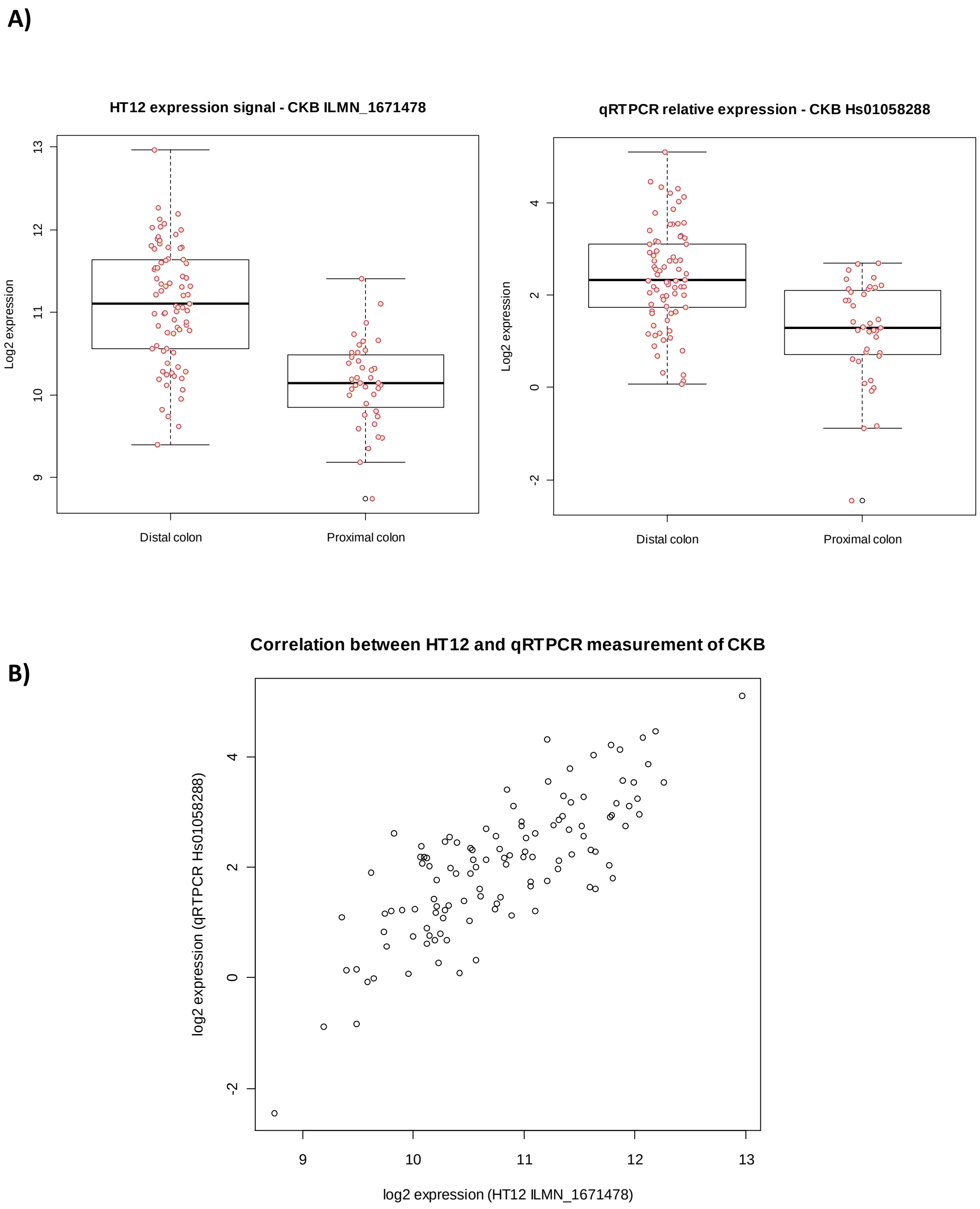


Figure 5.6. **A)** Boxplots of *CKB* expression as quantified by the Illumina HT12 microarray or qRT-PCR (unpaired t-test, *p*=3.6E-07). **B)** Scatterplot demonstrating the relationship between the expression of *CKB* as measured by the two different techniques (Spearman rho=0.759, *p*-value=< 2.2e-16).

A functional enrichment analysis indicated that genes that were more highly expressed in the proximal colon were enriched in genes involved in substrate metabolism processes, whereas genes that were more highly expressed in the distal colon were enriched for secreted proteins (Table 5.9).

A)

Category	**Term**	**FDR adjusted *q*-value**
KEGG_PATHWAY	Pentose and glucuronate interconversions	1.30E-05
KEGG_PATHWAY	Ascorbate and aldarate metabolism	1.40E-05
PIR_SUPERFAMILY	Glucuronosyl*trans*ferase	1.40E-05
KEGG_PATHWAY	Drug metabolism	1.70E-05
KEGG_PATHWAY	Metabolism of xenobiotics by cytochrome P450	2.20E-05
KEGG_PATHWAY	Retinol metabolism	2.90E-05
INTERPRO	UDP-glucuronosyl/UDP-glucosyl*trans*ferase	6.00E-05
KEGG_PATHWAY	Porphyrin and chlorophyll metabolism	7.10E-05
KEGG_PATHWAY	Androgen and estrogen metabolism	8.70E-05
KEGG_PATHWAY	Drug metabolism	1.10E-04
KEGG_PATHWAY	Starch and sucrose metabolism	1.10E-04
KEGG_PATHWAY	Steroid hormone biosynthesis	1.20E-04
GOTERM_MF_FAT	glucuronosyltransferase activity	3.60E-04
GOTERM_CC_FAT	extracellular space	4.60E-04
GOTERM_CC_FAT	extracellular region part	2.60E-03
UP_SEQ_FEATURE	signal peptide	2.80E-03
SP_PIR_KEYWORDS	signal	3.20E-03
GOTERM_CC_FAT	extracellular region	1.30E-02
SP_PIR_KEYWORDS	microsome	2.10E-02

B)

Category	**Term**	**FDR adjusted *q*-value**
SP_PIR_KEYWORDS	Secreted	5.2E-03

Table 5.9 Functional annotation of pathways, processes and GO terms that are over-represented in genes that were more highly expressed by >1.5 fold in A) proximal colon; B) distal colon. This enrichment analysis was performed using DAVID (as described in 2.6.3) using the tool's default *Homo sapiens* whole-genome background list.

5.4 Discussion

The examination of whole-genome gene expression profiles in relation to clinical variables thought to influence gene expression in the colon has detected gender- and anatomical site-specific expression differences. There were no detectable age- or cancer-related effects, but the power of these analyses are likely to be limited by the small numbers and skewed demographics of patients presenting for non IBD (inflammatory bowel disease) -related large bowel resections.

The large majority of genes that are more highly expressed in the colorectal mucosa of my male subjects are Y-linked. Rather unexpectedly, there is an X chromosome gene *CD99* that appears to be more highly expressed in males. It is a pseudoautosomal gene with a role in innate immunity, and its expression has been reported to be higher in the monocytes of males at baseline levels and after in vitro lipopolysaccharide stimulation (Lefevre *et al*, 2012), albeit in a different tissue type. The directionality of this reported difference is reassuring and may be of general interest, as gender is known to influence the severity and evolution of various inflammatory conditions. However, gender-bias is not a general feature of inflammatory bowel disease, except in the Asian population where male predominance in IBD is typically observed.

On the other hand, the majority of the female-biased genes in the colon are X-linked and are recognised to escape, at least partially, X-inactivation. *XIST* is the obvious exception, as it is a non-coding RNA gene that is the major effector of the X-inactivation process and hence only expressed on the inactive X in females. A number of these X-linked genes have functionally equivalent Y homologues, and the higher expression in females is likely to reflect the mechanism by which dosage compensation between males and females are achieved. For instance, *RPS4X* (ribosomal protein S4, X-linked) is more expressed in females, whereas *RPS4Y1* (ribosomal protein S4, Y-linked) is more expressed in males. Other genes similarly implicated are *EIF1AX, DDX3X, UTX, ZFX, NLGN4X,* and *PRKX*. The other X-linked genes such as *ARSD* (Carrel *et al*, 2005), *HDHD1A* (Yen *et al*, 1993), *TRAPPC2* (Mumm *et al*, 2001), *UBE1* (Carrel *et al*, 1996), *OFD1* (Carrel *et al*,

2005) have previously been reported as escaping X-inactivation in other tissue types. The chromosome 19 gene *NLRP2* is not known to have gender differences in expression. However, it has been reported to be reduced in axial spondyloarthropathy (Sharma *et al*, 2009), which has a marked male predominance, and the authors could not exclude a possible effect of gender on the level of transcript expression. This is interesting as firstly, this is in line with the differential expression of *NLRP2* favouring females in the colorectal mucosa, and secondly, *NLRP2* is a component of some inflammasomes (Church *et al*, 2008) with inhibitory effects on NF-kB and activating effects on caspase 1. This may reflect common aetiological pathways as colorectal cancer also sees a male predisposition and the connection between inflammation and colorectal tumorigenesis is well-recognised (as reviewed by Terzić *et al*, 2010). Overall, the gender-specific differential expression detected in the colorectal mucosa is largely consistent with known biological processes and published literature, providing confidence in the integrity of the samples, the microarray platform used and the processing of the data. It also demonstrates the importance of adjusting for gender in any differential gene expression analysis.

Over half of the site-specific differentially expressed genes detected in my samples have been previously identified to be differentially expressed, including *PITX2, MB, ETNK1*, *SLC23A3*, *IGFBP2*, *SLC20A1*, *MEP1B, ANPEP, OASL, FAM3B, NRIH4, OSTalpha, CCL13, CCL8, DEFA5, PRAC, CLDN8, HOXB13, SPON1, CAPN13, CKB, MUC17, TFF1, CRIP1, WFDC2, SPINK5, MUC12, PYY, PI3, GCG* and *S100P*. Although the fold differences for these genes are not always consistent with other published studies, the directionality of the differential expression are in accordance with the findings of other similar investigations (Birkenkamp-Demtroder *et al*, 2005; Glebov *et al*, 2003; LaPointe, 2008).

Five genes were selected for qRT-PCR validation, of which *PRAC*, *PITX2* and *CKB* have previously been described to have site-specific differences in transcript abundance. *PRAC* is known to be expressed only in human prostate, prostate cancer, rectum and the distal colon. Possible co-transcription with *HOXB13* and sequence analysis suggests a regulatory role in the nucleus (Liu *et al*, 2001). *PITX2* is responsible for the establishment of the left-right axis (Logan *et al*, 1998), asymmetrical development of visceral organs (Shiratori *et al*, 2006) and gut looping (Campione *et al*, 1999). It is overexpressed in colorectal cancer (Hirose *et al*, 2011) and has been shown to be induced by the Wnt/beta-catenin pathway and required for cell-type-specific proliferation (Kioussi *et al*, 2002). *CKB* is a cytosolic isoform of creatine kinase that is central in cellular energy homeostasis, and has been previously shown to promote EMT (Mooney *et al*, 2011). *L1TD1* and *OLFM4* are novel genes with detectable site-specific expression. Apart from providing technical validation, these two novel genes are interesting findings due to their known function in normal and diseased tissue. *L1TD1* codes for a stem-cell specific RNA-binding protein that has a role in the regulation of stemness. It has been shown to be abundantly expressed in undifferentiated human embryonic stem cells (Wong *et al*, 2011) and influences their self-renewal, acting downstream of pluripotent cell-specific transcription factors OCT4, SOX2 and NANOG (Narva *et al*, 2012). *OLFM4* is a marker for stem cells in the human intestine with restricted expression at the crypt base columnar cells (van der Flier *et al*, 2009). It is a glycoprotein that is selectively expressed in inflamed colorectal epithelium and secreted into the mucus in active IBD (Gersemann *et al*, 2012), as well as a candidate biomarker for adenomas and non-metastatic colorectal cancer (Besson *et al*, 2011). There is also evidence that *OLFM4* exerts an influence on the host defense against *H. pylori* infection by acting through NOD1/NOD2 mediated NF-kB activation (Liu *et al*, 2010). The site-specific expression of these two genes is suggestive of differences in stem cell dynamics and self-renewal regulatory properties between the proximal and distal colon, as well as the cellular response to inflammatory processes. This may in turn have an influence on transformation and the initiation of cancer, particularly as the cancer cell of origin is thought to originate from adult intestinal stem cells. Protein quantification and localisation within the large intestinal crypt of these site-specific expression differences would provide an additional degree of validation, and offer further insight into how these site-specific differences are relevant to the aetiology of tumourigenic pathways.

Apart from lending support to the notion that left and right-sided CRC tumours have distinctive aetiologies, these findings also suggest that regulatory mechanisms are distinctively different between the two segments. Hence, it is important to account for these potentially confounding differences when performing genotype-gene expression association analysis in the normal mucosa. Linear regression modelling conditional on the sampling site will serve to remove any confounding effects of site-specific expression, if by chance more of one allele is found in samples taken from a certain side of the colon. However, this may be an over-simplistic view as the majority of these normal mucosa samples are harvested from surgical specimens resected from patients with colorectal cancer/adenomas, and the site of the normal mucosa almost always represents the site of the cancer. There is a possibility that co-segregation of genotype and sampling site (representing the tumour site) are not arising by chance, i.e. left- and right-sided tumours have different genetic predispositions. GWAS to date have examined genetic predisposition to CRC cancer in relation to their location, but this categorisation was limited to the colon versus the rectum, with no further subdivision into the proximal or distal colon. For instance, the 11q23.1 locus confers risk that was greater for rectal than for colon cancer (Tenesa *et al*, 2008), and one might speculate that there could also be differential genetic risk within the colon. Also, there may be site-specific eQTL that may not be detected even when the modelling is conditioned for site, as any effect in one site could be masked by the lack of/opposing effect in the other site. Analysing the samples as two distinct tissue types (proximal colon vs distal colon) would be more ideal in this regard; however, this will significantly reduce the sample size and subsequently the power to detect eQTL. Due to the relatively small sample size (n=115) and the targeted nature of my analysis to the 25 colorectal risk variants only, it was decided not to analyse the different sites separately to maintain power. This will be of more importance in the longer term, when more samples are collected and added to this dataset, especially for future whole-genome eQTL studies that are already ongoing using this data.

Overall, the identification of factors that affect global expression in the tissue substrate demonstrates the importance of examining and adjusting for them in any genotype-gene expression analysis, as this will reduce the noise within the expression phenotype and improve detection. In addition, this analysis has identified some novel differentially expressed genes that could potentially inform other studies of colorectal cancer aetiology.

Chapter 6

Cis-eQTL analysis of low-penetrance common genetic variants associated with colorectal cancer predisposition

6.1 Introduction

At the point of the conception of this study, there was no strong evidence that colorectal cancer associated genetic loci exhibits eQTL effects, although two of them are in linkage disequilibrium with *cis*-eQTLs – rs7136702 was in moderate to strong LD with four SNPs previously associated with DIP2B expression in lymphoblastoid cell lines, and rs492536 was in moderate to strong LD with an eQTL for LAMA5 expression in the liver tissue (Houlston *et al*, 2010). This provides some functional evidence for these risk variants, as studies have shown that a substantial number of eQTLs are shared across diverse tissue types (Bullaughey *et al*, 2009; Zeller *et al*, 2010). Other studies, however, have indicated that eQTLs are likely to be cell- or tissue-specific (Cowley *et al*; 2009, Dimas *et al*; 2009) – this may explain the lack of eQTL associations seen so far with CRC susceptibility loci. Furthermore, it is yet unclear how risk alleles exert their causative effect, either directly within the target tissue or by modifying the cellular phenotypes of other cell types that in turn act upon the target tissue. Careful deliberation should therefore be given when defining the tissue/cell substrate for eQTL studies.

Considering that colorectal carcinomas are epithelial in origin, examination of eQTLs within the normal non-aberrant colonic mucosa and matched peripheral blood mononuclear cells would serve as a useful empirical starting point for the functional characterisation of CRC-associated loci in a systematic manner, with the hypothesis that they exert their effect on risk by influencing baseline gene expression. As the sample size of my study is relatively small (n=115), I will focus on the possibility of their role as *cis* determinants of gene expression, as polymorphic *cis*-acting variants often have a large effect on the expression level of the target gene and are easier to detect than trans-acting variants (as reviewed by Cheung *et al*, 2003). Initially, the expression of genes within 2Mb upstream or downstream of the tagging SNPs will be analysed in relation to the genotype of the tagging SNPs. Any eQTL association will be followed-up with fine association mapping, firstly, to identify candidate functional variants, and secondly, to identify with better precision where the regulatory variants are relative to the target genes. Case-control comparisons of these putative eQTL functional variants will then be performed, with the rationale that a variant that better explains both target gene expression and CRC risk will be more likely to represent the functional variant within a risk locus.

6.2 Methodological overview

6.2.1 Study subjects and biological material

To explore the effect of risk variants on gene expression in the normal colonic mucosa, 115 fresh normal colorectal mucosa were harvested immediately as described in 2.1 after surgical resection of colorectal adenocarcinomas (n=99), tubulo-villous (n=8) and villous (n=2) adenomas or non-neoplastic conditions (n=6). Matched PBMC were collected from 60 of the 115 subjects, of which 59 were used for gene expression analysis. The characteristics of the study subjects are detailed in Table 5.3 in Chapter 5.

6.2.2 RNA extraction and gene expression quantification of *cis*-genes

RNA was isolated from the tissue samples as described in 2.3.1. The expression of the *cis*-genes within a 2Mb radius of each individual risk locus was derived from whole-transcriptome gene expression profiling of the normal mucosa and PBMC as described in 2.3.6.

6.2.3 Genotyping of CRC risk loci

Genotypes were obtained by hybridising genomic DNA extracted from EDTA-venous blood on the HumanOmni5M-4v1_B BeadChip Arrays (Illumina, USA), which includes ~5 million SNP markers. Genotypes for 3 SNPs (rs4813802, rs10911251 and rs3824999) were not available on the array platform and were imputed using IMPUTE2 after phasing with SHAPEIT.[1] The 1000 genomes panel (Pilot1 Version3 release) was used as a reference. Post-imputation SNPs with an info value of <0.3 were excluded. All variants were annotated and presented according to the human reference sequence build GRCh37.p12 (hg19).

[1] Imputation performed by Maria Timofeeva, MRC Human Genetics Unit

6.2.4 Local eQTL analysis and regional fine-mapping

eQTL analysis between genetic variants and level of gene expression was performed using linear regression as implemented in the R package "MatrixEQTL" using an additive genetic model.[2] The analysis of normal mucosa was adjusted for age, gender and anatomical sampling site, whereas the analysis of PBMC was adjusted for age and gender only. For X-linked SNPs, male hemizygotes were treated as homozygotes. For each GWAS SNP, a distance of 2Mb upstream and downstream was used as a cut-off for *cis*-genes. A minimum significance level of nominal p=0.01 was considered relevant for reporting. To account for multiple comparisons, a Bonferroni correction was applied, taking into account the number of genes tested within each 4Mb region. The Bonferroni correction method was used based on the assumption that each individual test was independent of each other, and was selected over the Benjamini-Hochberg FDR procedure as there were relatively small numbers of *cis*-genes tested for each locus.

For each eQTL identified, fine-mapping of the region in linkage disequilibrium with the tagging SNP was performed in relation to the expression of the target gene. Regional visualisation of the fine-association mapping was performed using the web tool LocusZoom (Pruim *et al*, 2010; URL6.1) to produce Manhattan plots that display the strength of genetic association (-$\log_{10}$ p-value) with target gene expression versus chromosomal position. Each dot represents a genotyped or imputed SNP, and dot colours signify the degree of pairwise correlation (r^2) with the tagging SNP, as presented in the colour key. Grey dots depict SNPs for which r^2 values are unknown.

6.2.5 Case-control study of candidate functional SNPs

Candidate functional SNPs identified from eQTL fine-association mapping were evaluated for their association with CRC in a case-control study (Scotland Phase 1; cases=939, controls=945; males=965, females=919).[3] Genotypes were obtained by

[2] Analysis performed in collaboration with Maria Timofeeva, MRC Human Genetics Unit

[3] Case-control population samples were previously collected and genotyped by the Colon Cancer Genetics Group

hybridising genomic DNA extracted from EDTA-venous blood on the Illumina 300K or Illumina 240K BeadChip Arrays (Illumina, USA), and SNPs not available on these arrays were similarly imputed as described in 6.2.3. Similar to the cis-eQTL analysis, male hemizygotes were treated as homozygotes for X-linked SNPs.

6.2.6 Technical and biological validation of the Xp22.2 eQTL locus

Technical validation of the association between rs5934683 genotype and *SHROOM2* expression was performed with qRT-PCR using the Taqman® Gene Expression assay (ABI) (Table 6.1). The correlation between HT12 microarray expression and qRT-PCR was tested statistically with the Spearman correlation test. The association between rs5934683 genotype and *SHROOM2* expression as quantified by qRT-PCR was analysed in R, using linear regression modelling corrected for age, gender and anatomical sampling site.

	ILLUMINA HT12 MICROARRAY	**TAQMAN® GENE EXPRESSION ASSAY**
Assay/Probe ID	ILMN_1681777	Hs01113636_m1
Context Sequence	CCTGTCAGTTCCCCTGTTTGCCTCTG AAACGTCTGGTTAGTGGGGACCCAA	CTCCCGGTGATCGGCAATCACTGCT
Transcripts Detected	SHROOM2-001 (Exon 10)	SHROOM2-001 (exon 10) SHROOM2-201 (exon 6) SHROOM2-002 (exon 6)
Normalisation	Quantile normalisation	Reference genes *TBP*, *EIF2B1* and *RPL30*

Table 6.1 A technical comparison between the two methods used for detection of *SHROOM2* expression.

A further level of biological replication was introduced at the level of normal mucosa sampling and RNA extraction. In a subset of 37 patients (males=15, females=22), 2 further RNA extracts were prepared from the normal mucosa harvested at the same time as the first extract but with spatial variation within the same colonic site. *SHROOM2* was quantified with qRT-PCR, and a nested one-way ANOVA was used to statistically account for the multiple levels of replication.

6.3 RESULTS

6.3.1 Local eQTL associations of CRC risk loci in normal colorectal mucosa and PBMC

In the normal colorectal mucosa, 15 SNP-gene expression associations in ten risk loci were identified (nominal *p*-value <0.01) for genes within 2Mb radius of the CRC risk variants (Table 6.2). The more stringent Bonferroni correction was performed to reduce the number of false positives as any associations will require follow-up with validation studies. Five of these associations were significant (adj. *p* <0.05) after Bonferroni correction. In PBMCs, 13 SNP-gene expression associations in seven risk loci were identified, of which five were significant (*p* <0.05) after Bonferroni correction (Table 6.3). One eQTL was present in both tissue types (rs7136702-CERS5), but the association in the normal mucosa did not survive multiple correction testing. Several other SNPs had local eQTL effects in both tissue types, but the genes that were associated did not overlap and the majority of them were only nominally significant in at least one of the tissue types. Two variants, rs11169552 (12q13.12) and rs16892766 (8q23.3) had local eQTL associations in both tissue types that survived multiple testing. rs11169552 is associated with the expression of *SPATS2* in the normal mucosa (adj. *p*=0.033) and expression of *LIMA1* in PBMC (adj. *p*=0.046), whereas rs16892766 is associated with the expression of *UTP23* in the normal mucosa (adj. *p*=0.008) and expression of *MED30* in PBMC (adj. *p*=0.033).

SNP	Locus	Gene	Description	*p-val*	Adj. *p*	Beta
rs3802842	11q23.1	*COLCA2*	colorectal cancer associated 2	6.85e-14	1.92e-12	-0.16
rs3802842	11q23.1	*COLCA1*	colorectal cancer associated 1	2.16e-11	6.05e-10	-0.11
rs5934683	Xp22.2	*SHROOM2*	shroom family member 2	4.17e-10	5.00e-09	-0.14
rs11169552	12q13.12	*SPATS2*	spermatogenesis associated, serine-rich 2	4.11e-04	3.25e-02	-0.06
rs16892766	8q23.3	*UTP23*	small subunit (SSU) processome component, homolog (yeast)	8.75e-04	7.88e-03	-0.06
rs4925386	20q13.33	*OSBPL2*	oxysterol binding protein-like 2	1.73e-03	1.21e-01	0.09
rs6687758	1q41	*HLX*	H2.0-like homeobox	2.13e-03	6.17e-02	0.04
rs7136702	12q13.12	*CERS5*	ceramide synthase 5	2.23e-03	1.85e-01	-0.06
rs3217810	12p13.32	*TEAD4*	TEA domain family member 4	2.79e-03	1.12e-01	-0.03
rs11169552	12q13.12	*SMARCD1*	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily d, member 1	2.98e-03	2.35e-01	-0.08
rs16969681	15q13.3	*NOP10*	NOP10 ribonucleoprotein	5.31e-03	1.59e-01	0.13
rs4925386	20q13.33	*KCNQ2*	potassium channel, voltage gated KQT-like subfamily Q, member 2	7.16e-03	5.01e-01	-0.03
rs16969681	15q13.3	*FAN1*	FANCD2/FANCI-associated nuclease 1	8.22e-03	2.47e-01	-0.12
rs1321311	6p21.2	*FGD2*	FYVE, RhoGEF and PH domain containing 2	9.23e-03	4.25e-01	0.15
rs7136702	12q13.12	*PRPH*	peripherin	9.70e-03	8.05e-01	-0.15

Table 6.2 CRC risk SNPs that show an association (nominal *p*-value<0.01) with the expression of *cis*-genes in normal colorectal mucosa (n=115). Associations that are significant (adj. *p*<0.05) after Bonferroni correction are highlighted in grey.

SNP	Locus	Gene	Description	*p*-val	Adj. *p*	Beta
rs7136702	12q13.12	*CERS5*	ceramide synthase 5	1.0e-05	7.66e-04	-0.12
rs11169552	12q13.12	*LIMA1*	LIM domain and actin binding 1	6.5e-04	4.63e-02	-0.08
rs961253	20p12.3	*RP11-19D2.2*	lincRNA	6.7e-04	1.47e-02	-0.05
rs1321311	6p21.2	*MDGA1*	MAM domain containing glycosylphosphatidylinositol anchor 1	1.1e-03	4.86e-02	0.06
rs1321311	6p21.2	*CMTR1*	cap methyltransferase 1	2.8e-03	1.28e-01	0.12
rs16892766	8q23.3	*MED30*	mediator complex subunit 30	4.2e-03	3.32e-02	0.12
rs1321311	6p21.2	*MAPK14*	mitogen-activated protein kinase 14	6.7e-03	3.07e-01	-0.10
rs11169552	12q13.12	*ATF1*	activating transcription factor 1	6.8e-03	4.79e-01	-0.04
rs7136702	12q13.12	*SPATS2*	spermatogenesis associated, serine-rich 2	9.5e-03	7.10e-01	0.05
rs11169552	12q13.12	*DIP2B*	DIP2 disco-interacting protein 2 homolog B (Drosophila)	9.7e-03	6.92e-01	0.06
rs3824999	11q13.4	*GDPD5*	glycerophosphodiester phosphodiesterase domain containing 5	5.2e-03	2.36e-01	-0.14
rs6691170	1q41	*DUSP10*	dual specificity phosphatase 10	7.8e-03	2.03e-01	-0.03
rs6691170	1q41	*MARC1*	mitochondrial amidoxime reducing component 1	9.1e-03	2.35e-01	0.28

Table 6.3 CRC risk SNPs that show an association (nominal *P* value<0.01) with the expression of *cis*-genes in PBMC (n=59). Associations that are significant (adj. *p*<0.05) after Bonferroni correction are highlighted in grey.

6.3.2 Identification of putative functional variants underlying individual eQTL associations in the normal mucosa

In the normal mucosa, the strongest association was seen with rs3802842 which tags the locus 11q23.1. This SNP was found to be highly associated with the two genes that it is intronic to: *COLCA2* (adj. *p*-val=1.92e-12) and *COLCA1* (adj. *p*-val=6.05e-10), two uncharacterised genes that appear to be co-regulated and transcribed from opposite strands. These 11q23 eQTL associations have recently been published by two separate groups (Closa *et al*, 2014; Peltekova *et al*, 2014), providing independent replication and validation to our findings. The 11q23.1 locus corresponds to a 150kb region of LD, and fine association mapping of the region showed that five SNPs that are in high LD with rs3802842 were more significantly associated with the expression of COLCA2 (Figure 6.1), four of which were also more significantly associated with the expression of COLCA1 (Figure 6.2). This suggests that they may be better functional candidates than the tagging SNP. In a CRC case-control study of 939 cases and 945 controls, only one of these SNPs rs11213801 showed a marginally better association with CRC risk (Table 6.4). Further genotyping and analysis of the variation within this locus was taken forward by fellow PhD student Claire Smillie.

The second locus that exhibited eQTL properties was rs59364683 at Xp22.2. This SNP is intronic to a putative *GPR143* transcript but an association with this gene was not observed (nominal *p*=0.083). Instead, a strong association was observed with the expression of neighbouring gene *SHROOM2* (adj. *p*=5.0e-09), which lies approximately 3kb downstream from the locus tagging SNP. Indeed, this *cis*-eQTL association was detectable even in a preliminary analysis of *SHROOM2* expression in an early subset of these normal mucosa samples (n=42, nominal *p*=1.3e-07), accounting for 55% of the variation in *SHROOM2* expression (as reported in Dunlop *et al*, 2012). The *SHROOM2* association was also replicated in another study recently (Closa *et al*, 2014), but the authors also predicted an association with *GPR143* which was not observed in my samples. Fine-mapping of this eQTL locus revealed four SNPs within the first intron of *SHROOM2* that are more highly associated with *SHROOM2* expression (Figure 6.3). All four SNPs were also more significantly associated with CRC risk with higher odds ratio than the tagging SNP (Table 6.5), which is suggestive of a functional role.

The two other CRC risk SNPs that exhibited local eQTL effects, albeit much weaker, were rs11169552 (12q13.12) and rs16892766 (8q23.3). rs11169552 is an intergenic SNP that has been recently shown to be an eQTL for neighbouring gene *DIP2B* (Closa *et al*, 2014) but this was not seen in my dataset (nominal *p*= 0.14 and 0.86, two expression probes present). Fine-mapping of this region for both *DIP2B* probes did not reveal any eQTL associations in the 500kb region in LD with the tagging SNP (Figure 6.4). rs11169552 is, however, associated with expression of *SPATS2* which is approximately 1.2Mb upstream (adj. *p*=0.033). It is the best eQTL in this region for SPATS2, and is in high LD with four other intronic SNPs within neighbouring gene *ATF1* (Figure 6.5). These four SNPs appear to be in perfect LD with one another, and likely represent a single genetic signal.

rs16892766 (8q23.3) is an intergenic variant that appears to be an eQTL locus for nearby gene *UTP23* (adj. *p*=0.008). Fine-mapping of the region revealed 11 other SNPs in LD that are more significantly associated with expression of *UTP23* (Figure 6.6); the majority of them are intronic variants within *EIF3H* and *UTP23*, with one missense variant (rs16888728) that is predicted by SIFT and PolyPhen to be a tolerated/benign variant. However, none of these SNPs showed an association with CRC risk in the case-control comparison (Table 6.6).

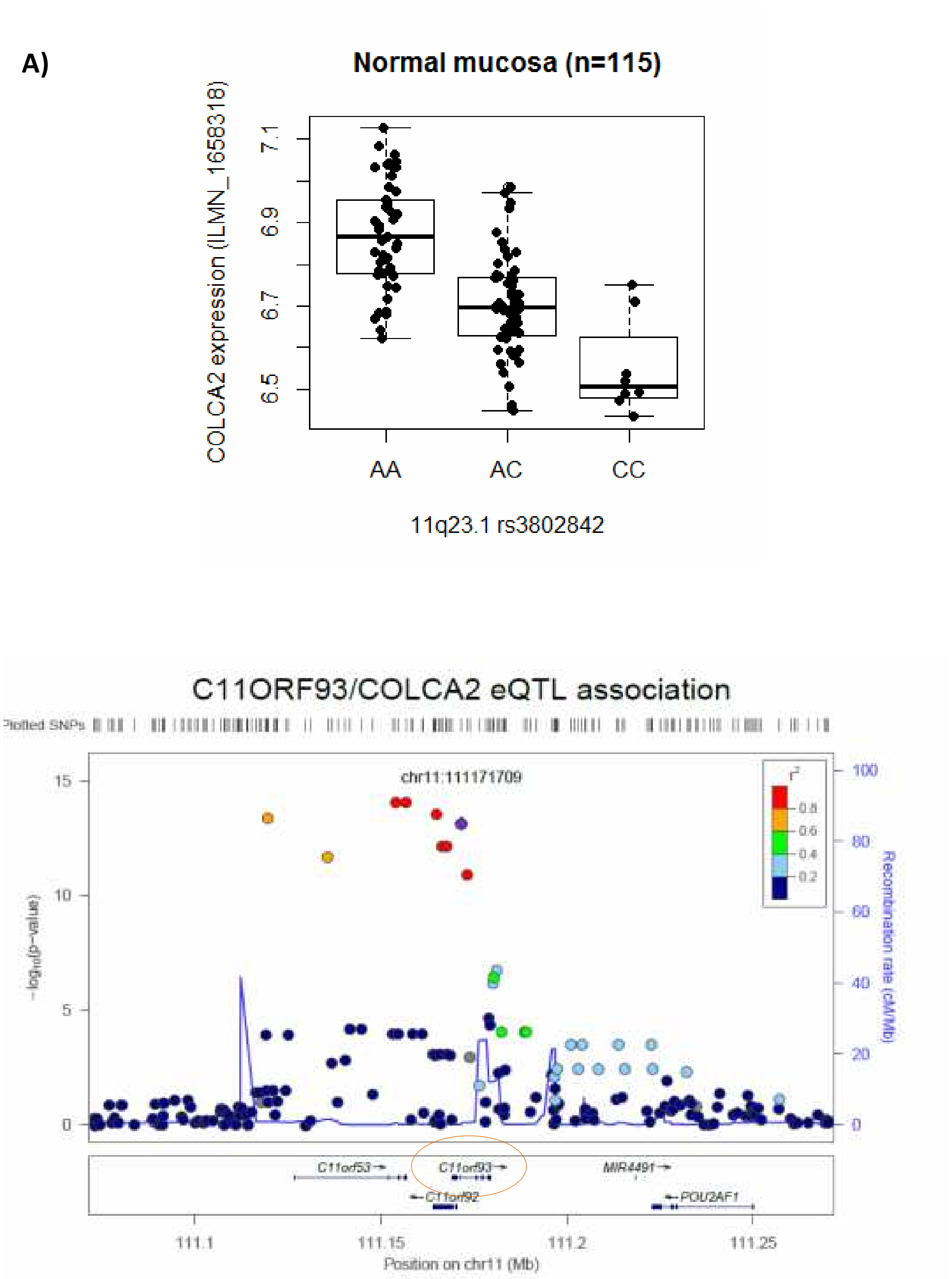


Figure 6.1 A) Boxplot showing the rs3802842 genotype association with *COLCA2* (also known as C11orf93) expression in normal colorectal mucosa. B) Fine association mapping for *COLCA2* expression detailing a 150kb region in LD with rs3802842 (purple).

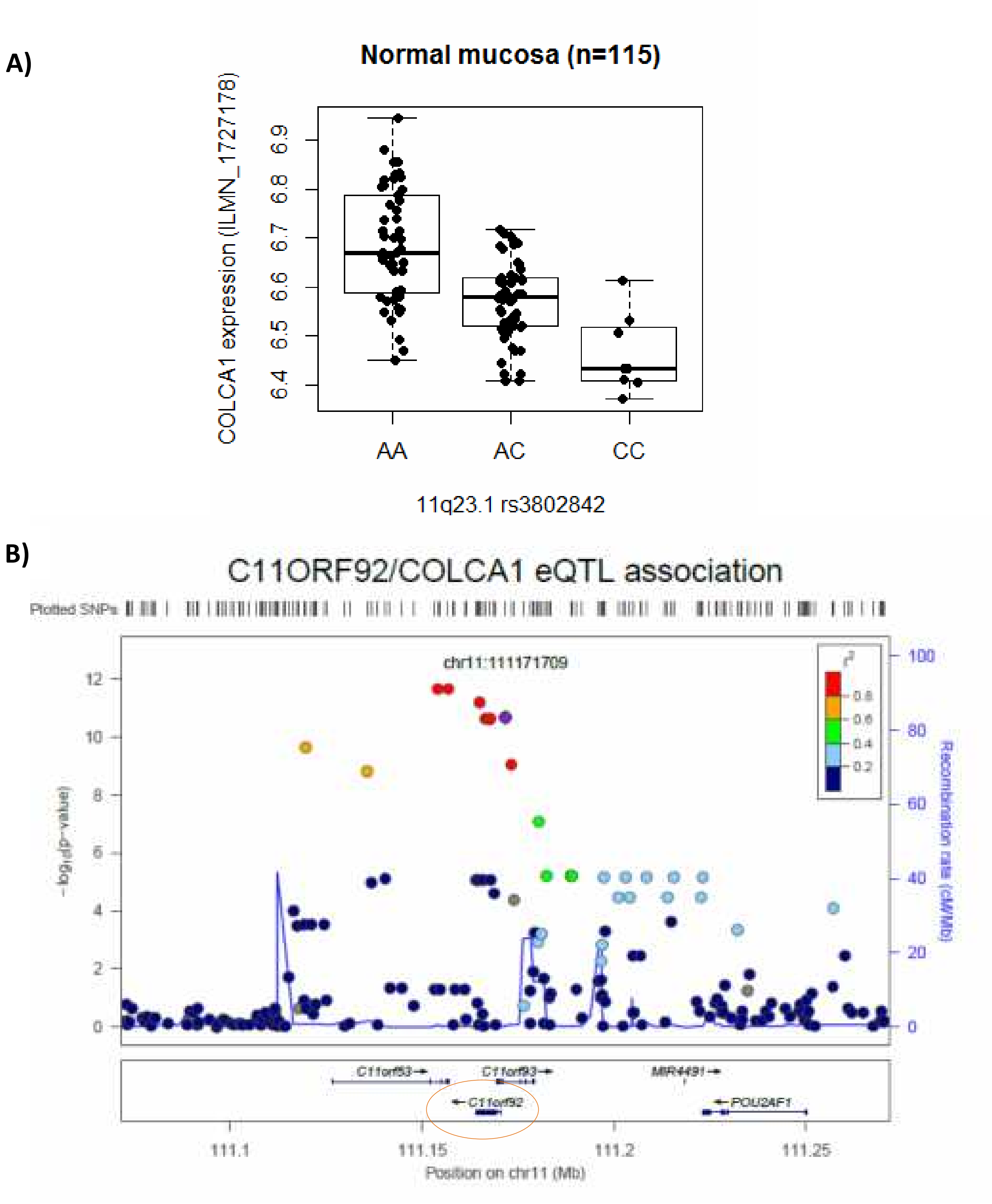


Figure 6.2 A) Boxplot showing the rs3802842 genotype association with *COLCA1*(also known as C11orf92) expression in normal colorectal mucosa. B) Fine association mapping for *COLCA1* expression detailing a 150kb in LD with rs3802842 SNP (purple).

A)

			eQTL		Case-control	
SNP	**SNP position**	**Predicted function**	***p*-value**	**beta**	***p*-value**	**OR**
rs3087967	chr11:111156836	*C11orf53* 3' UTR variant	8.53e-15	-0.165	9.22e-03	1.20
rs7130173	chr11:111154072	*C11orf53* intron variant	8.53e-15	-0.165	1.36e-02	1.20
rs7103178	chr11:111165009	*COLCA1* 3' UTR variant	2.58e-14	-0.159	8.35e-03	1.20
rs11213801	chr11:111119694	Intergenic	3.86e-14	-0.156	4.50e-03	1.24
rs3802840	chr11:111171646	*COLCA1* and *COLCA2* intron variant	6.85e-14	-0.161	6.02e-03	1.21
rs3802842	chr11:111171709	*COLCA1* and *COLCA2* intron variant	6.85e-14	-0.161	6.02e-03	1.21

B)

			eQTL		Case-control	
SNP	**SNP position**	**Predicted function**	***P*-value**	**beta**	***p*-value**	**OR**
rs3087967	chr11:111156836	*C11orf53* 3' UTR variant	2.35e-12	-0.116	9.22e-03	1.12
rs7130173	chr11:111154072	*C11orf53* intron	2.35e-12	-0.116	1.36e-02	1.19
rs7103178	chr11:111165009	*COLCA1* 3' UTR variant	6.67e-12	-0.112	8.35e-03	1.20
rs3802840	chr11:111171646	*COLCA1* and *COLCA2* intron variant	2.16e-11	-0.112	6.02e-03	1.21
rs3802842	chr11:111171709	*COLCA1* and *COLCA2* intron variant	2.16e-11	-0.112	6.02e-03	1.21

Table 6.4 Variants that are more significantly associated with normal mucosa A) *COLCA2,* and B) *COLCA1* expression than the tagging SNP rs3802842, listed in order of their eQTL significance values. The tagging SNP is included and highlighted for reference. *p*-values and effect sizes for their eQTL association and CRC risk (case-control comparison) are presented.

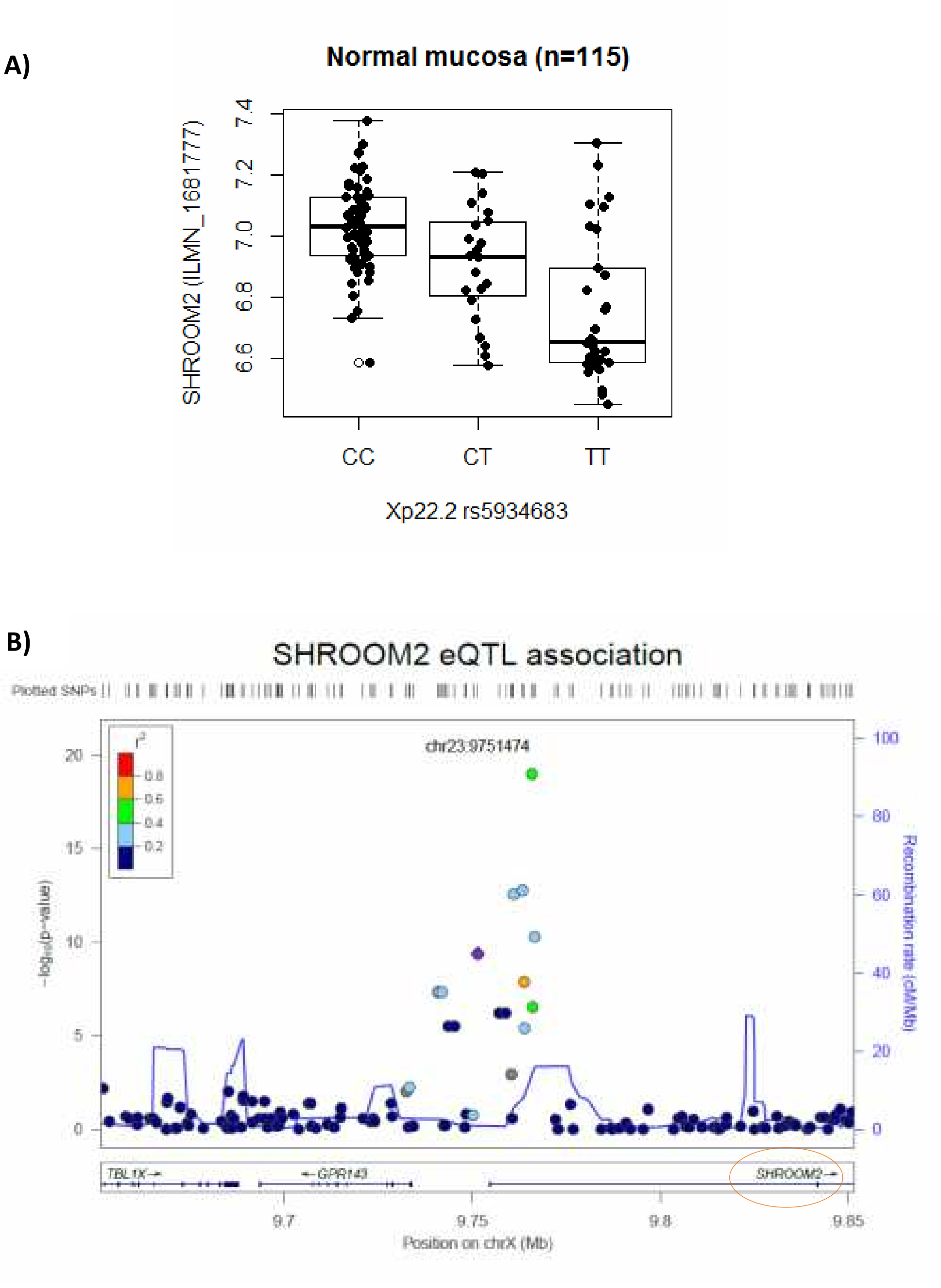


Figure 6.3 A) Boxplot showing the rs5934683 genotype association with *SHROOM2* expression in normal colorectal mucosa. B) Fine association mapping for *SHROOM2* expression detailing a 50kb region in LD with rs5934683 (purple).

			eQTL		Case-control	
SNP	**SNP position**	**Predicted function**	***p*-value**	**beta**	***p*-value**	**OR**
rs5934685	chrX:9766019	*SHROOM2* intron 1 variant	1.04e-19	-0.206	1.65e-02	1.296
rs2521664	chrX:9763429	*SHROOM2* intron 1 variant	1.62e-13	-0.159	3.52e-02	1.201
rs2521663	chrX:9761062	*SHROOM2* intron 1 variant	2.51e-13	-0.158	8.14e-02	1.158
rs4830657	chrX:9766725	*SHROOM2* intron 1 variant	5.28e-11	-0.147	6.21e-02	1.185
rs5934683	chrX:9751474	*GPR143* intron variant	4.17e-10	-0.138	5.50e-01	1.048

Table 6.5. Variants that are more significantly associated with normal mucosa *SHROOM2* expression than the tagging SNP rs5934683, listed in order of their eQTL significance values. The tagging SNP is included and highlighted for reference. *p*-values and effect sizes for their eQTL association and CRC risk (case-control comparison) are presented.

A)

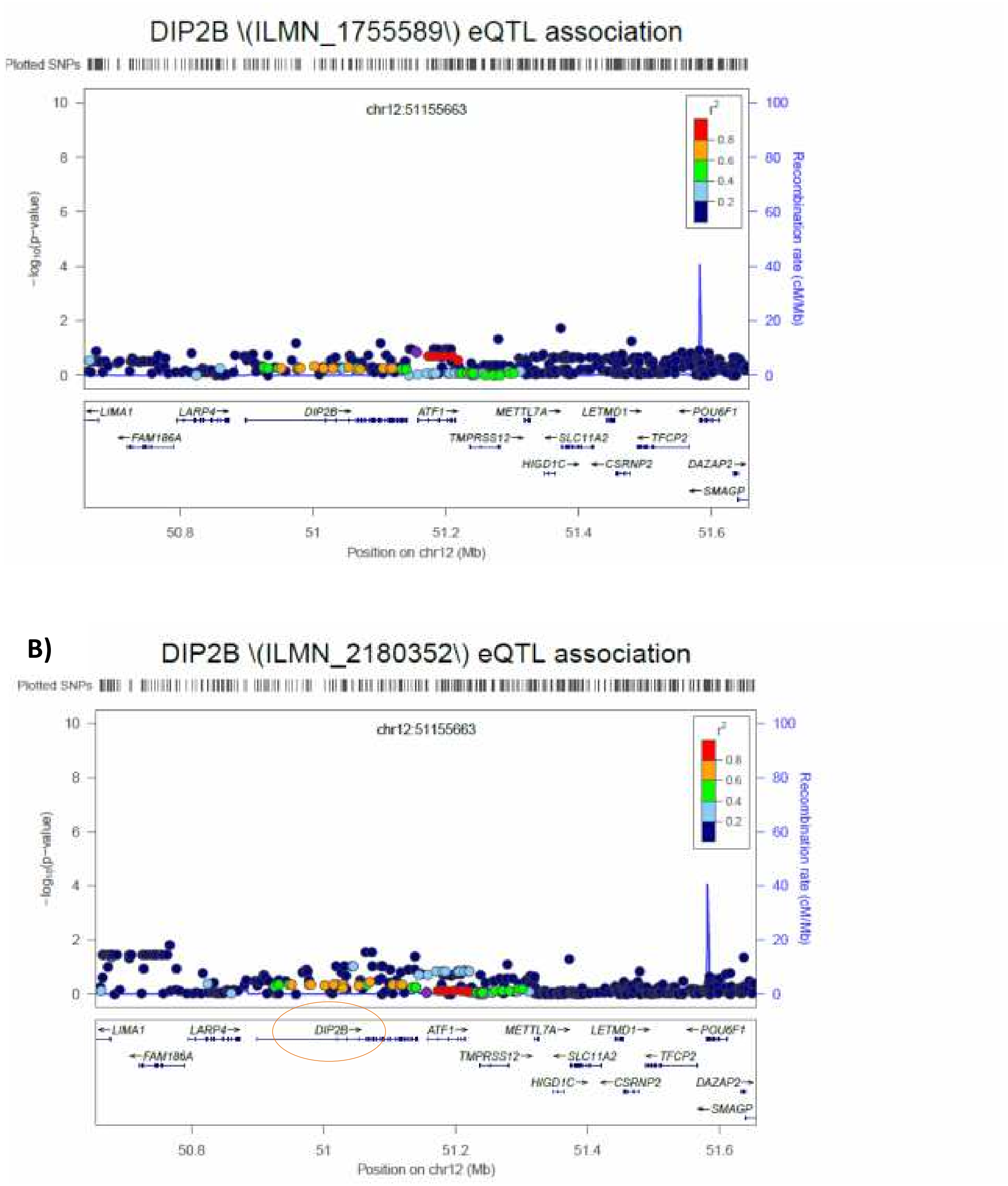


Figure 6.4. Fine association mapping for normal colorectal mucosa *DIP2B* expression detailing a 500kb region in LD with rs11169552 (purple). *DIP2B* expression was detected by two individual probes A) ILMN_1755589 and B) ILMN_2180352.

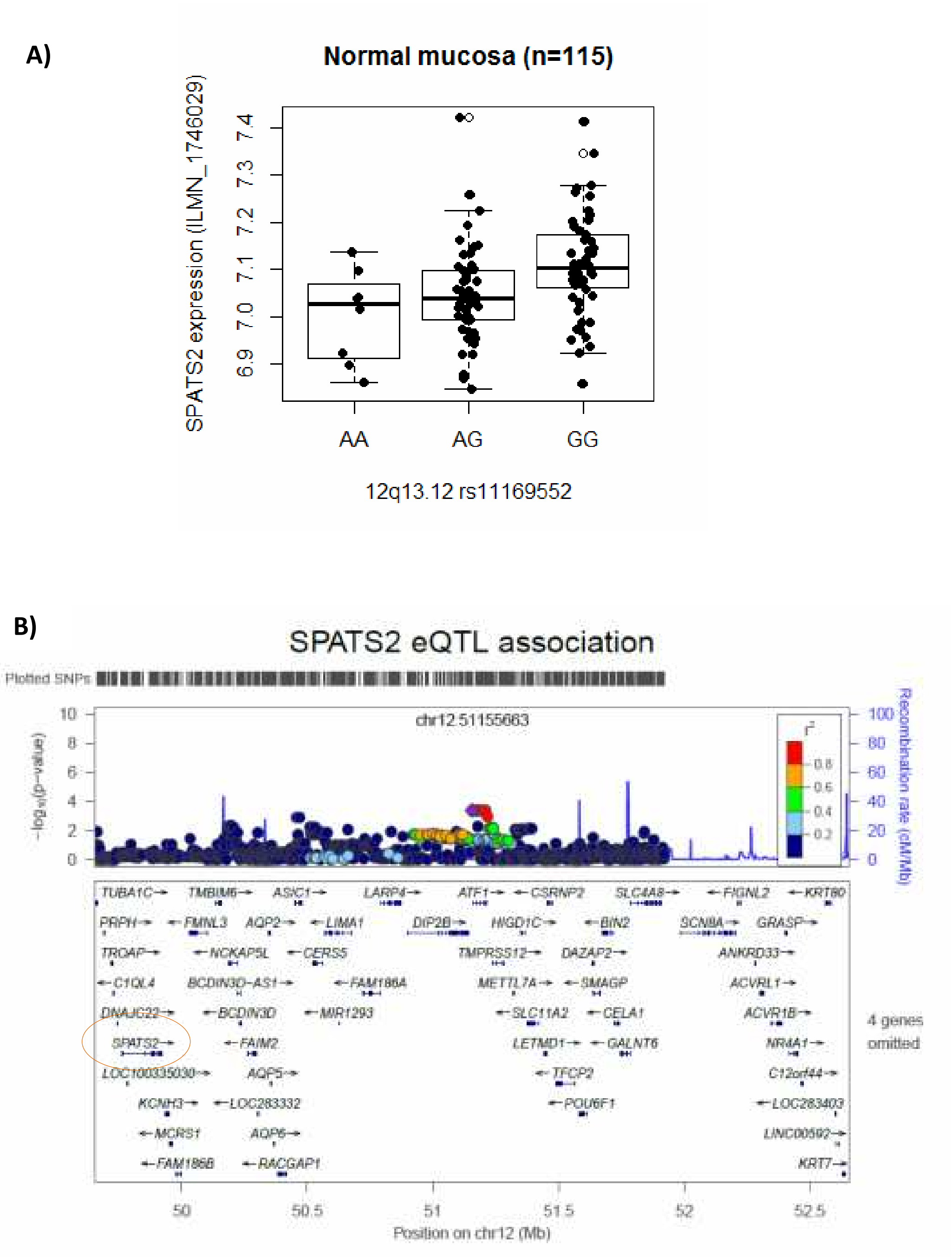


Figure 6.5 A) Boxplot showing the rs11169552 genotype association with *SPATS2* expression in the normal colorectal mucosa. B) Manhattan plot demonstrating *SPATS2* location in relation to the peak of the eQTL association. The 4 omitted genes are *PRPF40B, SMARCD1, GPD1, COX14.*

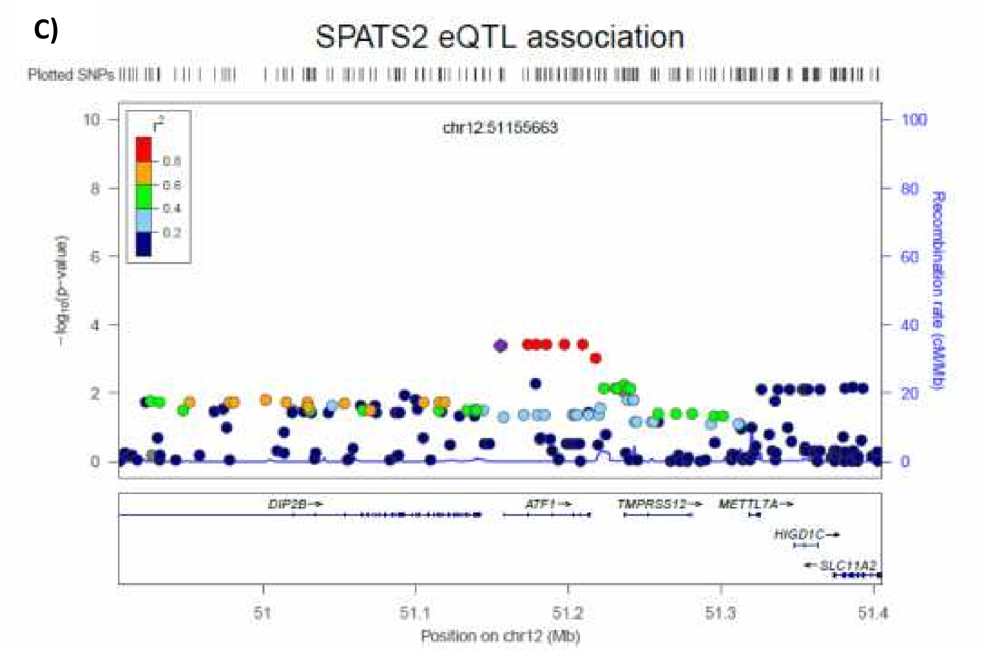


Figure 6.5 C) Fine association mapping for *SPATS2* expression detailing a 500kb region surrounding the rs11169552 SNP (purple).

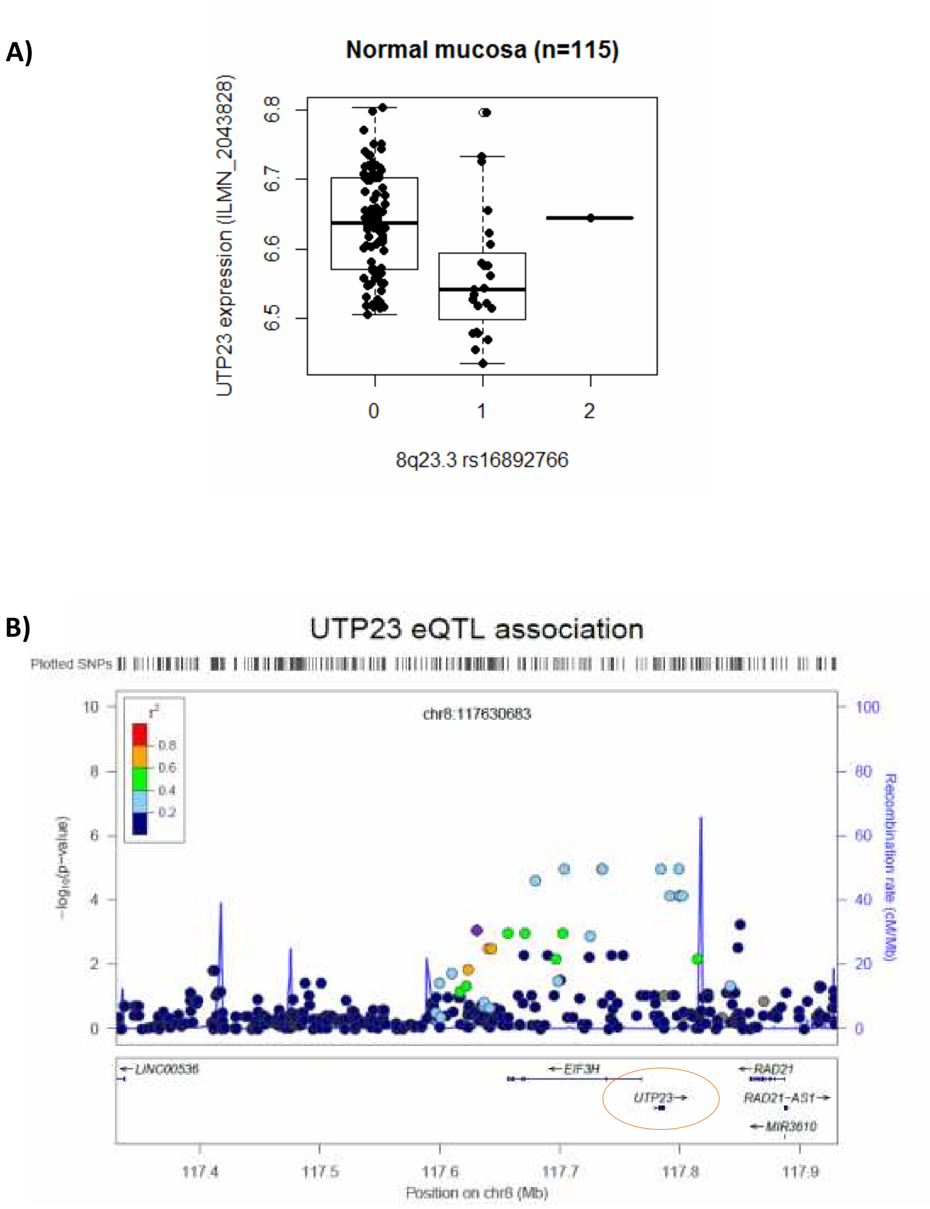


Figure 6.6 A) Boxplot showing the rs11169552 genotype association with *UTP23* expression in normal colorectal mucosa. B) Fine association mapping for *UTP23* expression detailing a 200kb region in LD with rs16892766 (purple).

			eQTL		Case-control	
SNP	**SNP position**	**Predicted function**	***p*-val**	**beta**	***p*-val**	**OR**
rs16892766	chr8:117630683	Intergenic	8.75e-04	-0.060	0.074	1.22
rs28668628	chr8:117679601	*EIF2H* intron variant	2.36e-05	-0.086	0.249	1.13
rs7823271	chr8:117703509	*EIF2H* intron variant	1.09e-05	-0.087	0.288	1.12
rs16888695	chr8:117735099	*EIF2H* intron variant	1.09e-05	-0.087	0.286	1.12
rs16888699	chr8:117735209	*EIF2H* intron variant	1.09e-05	-0.087	0.296	1.12
rs16888728	chr8:117783975	*UTP23* missense variant	1.09e-05	-0.087	0.292	1.12
rs979867	chr8:117791502	*UTP23* intron variant	7.30e-05	-0.081	0.269	1.12
rs1867840	chr8:117799012	*UTP23* 3' UTR variant	1.09e-05	-0.087	0.257	1.13
rs7014328	chr8:117799487	*UTP23* intron variant	7.30e-05	-0.081	0.265	1.12
rs200798730	chr8:117799586	*UTP23* intron variant	7.30e-05	-0.081	0.263	1.13
rs7014359	chr8:117799587	*UTP23* intron variant	7.30e-05	-0.081	0.280	1.12
rs6983626	chr8:117802148	*UTP23* intron variant	7.30e-05	-0.081	0.288	1.12

Table 6.6 Variants that are more significantly associated with normal mucosa *UTP23* expression than the tagging SNP rs16892766 are shown in the table, listed in order of their chromosomal positions. The tagging SNP is included and highlighted for reference. *p*-values and effect sizes for their eQTL association and CRC risk (case-control comparison) are presented.

6.3.3 Identification of putative functional variants underlying individual eQTL associations in PBMC

In PBMC, the strongest association was observed with rs7136702 (12q13.12) and expression of *CERS5* (adj. *p*=7.66e-04) that lies 350kb away (Figure 6.7). As alluded to before, this association was also observed in the normal mucosa but it was weaker and did not survive multiple testing correction (nominal *p* =2.23e-03, adj. *p*=0.185). Fine-mapping of the region revealed a 700kb region that is in LD with rs7136702, with 73 variants showing better association to *CERS5* expression. The peak of this association is striking, with the best eQTL rs10747573 (chr12:50633839) showing an association *p*-value = 4.80e-13. It is approximately 300kb closer to *CERS5* than the tagging SNP, and resides within a cluster of highly associated SNPs intronic to the upstream neighbouring gene *LIMA1*. However, none of the SNPs within this cluster were associated with CRC risk (Table 6.7). Examination of the wider LD block tagged by rs7136702 shows that the majority of these SNPs are intronic variants of the genes within this LD block, with a few synonymous variants and missense variants that are mostly predicted to be benign/tolerated. Of note, two missense variants within *FAM186A*, rs12303082 (chr12:50754563) and rs6580741 (chr12:50727706) are predicted by Polyphen to be probably damaging and possibly damaging, respectively. However, none of these are better candidates in predicting CRC risk. On the other hand, there are 5 other variants within this region that appear to be more significantly associated with CRC risk, the best candidate being *CERS5* intron variant rs7398567 (*p*=0.010). Taken altogether, this evidence is suggestive of *CERS5* being a candidate gene in CRC common predisposition.

For the genotype-gene expression associations rs11169552-*LIMA1*, rs961253-*RP11-19D2.2* and rs1321311-*MDGA1*, fine-association mapping did not show any other putative functional candidates that are better associated with the target gene expression (Figures 6.8 - 6.10). For rs16892766, there was one variant within *EIF3H* intron (rs7825662) that showed a better association with *MED30* expression (Figure 6.11), but it was not a significant predictor of CRC risk (Table 6.8). rs11169552 and rs16892766 are also eQTLs in the normal mucosa influencing different genes (*SPATS2* and *UTP23* respectively) as described previously.

Of these eight eQTL loci in the normal mucosa and PBMC, the two associations that stood out (11q23.3 and Xp22.2) were selected to be validated technically and functionally. Due to the collaborative nature of this project, the 11q23.3 locus and *COLCA1/COLCA2* expression was investigated by Claire Smillie and further data on this locus will not be presented in this thesis.

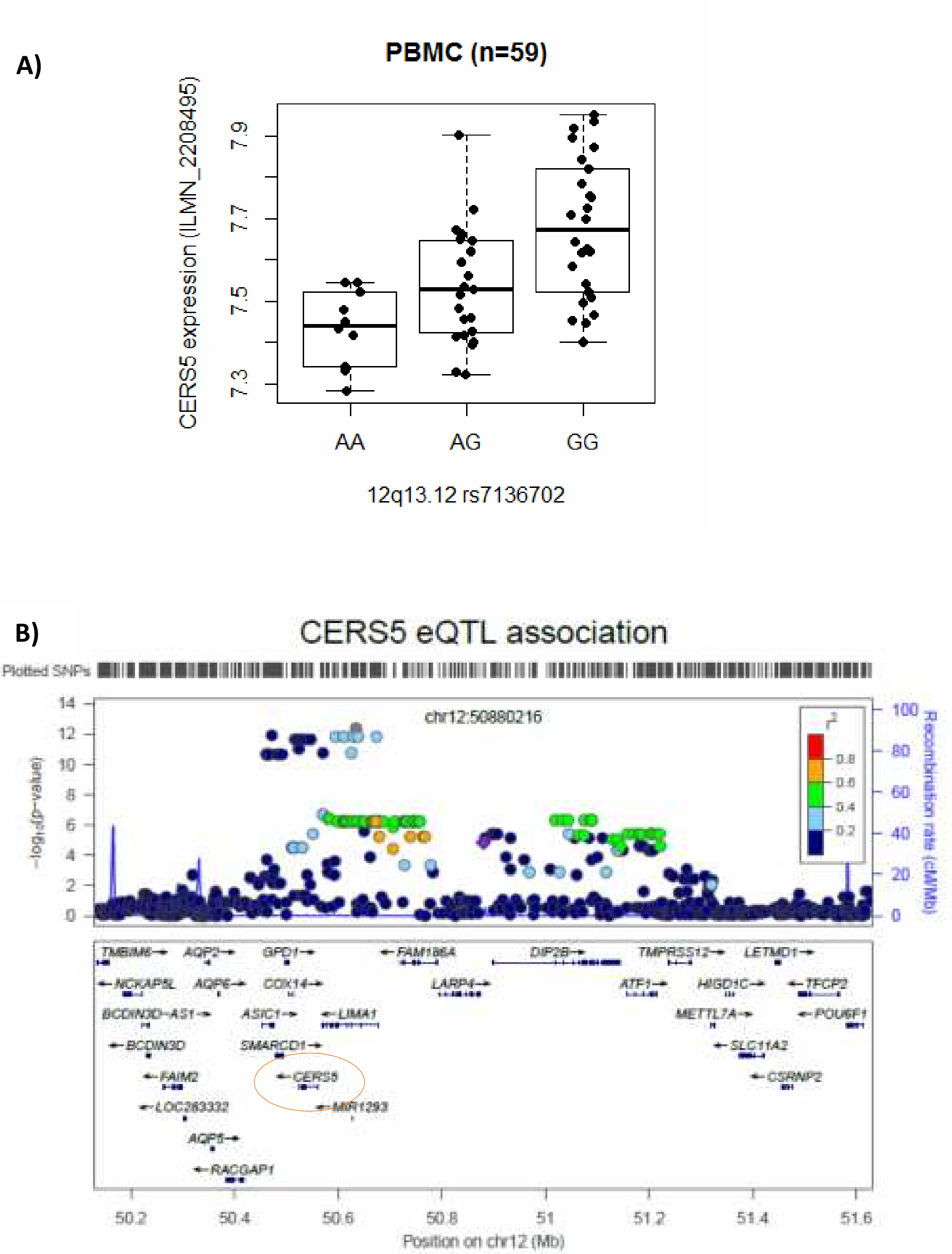


Figure 6.7 A) Boxplot showing the rs7136702 genotype association with *CERS5* expression in PBMC. B) Fine association mapping for *CERS5* expression detailing a 1Mb region in LD with rs7136702 (purple).

			eQTL		Case-control	
SNP	**SNP position**	**Predicted function**	***p*-val**	**beta**	***p*-val**	**OR**
rs7398567	chr12:50551158	*CERS5* intron variant	2.56e-06	-0.146	0.010	1.19
rs3184122	chr12:50570127	*LIMA1* 3'UTR variant	1.27e-07	-0.160	0.026	1.16
rs9364	chr12:50570519	*LIMA1* 3'UTR variant	1.27e-07	-0.160	0.025	1.16
rs7315690	chr12:50581490	*RP3-405J10.3-001* non coding transcript exon variant	2.41e-07	-0.157	0.058	1.14
rs7138420	chr12:50583150	*RP3-405J10.3-001* non coding transcript exon variant	2.41e-07	-0.157	0.065	1.13
rs2302900	chr12:50599709	*LIMA1* intron variant	4.01e-07	-0.152	0.073	1.13
rs12367872	chr12:50607834	*LIMA1* intron variant	4.01e-07	-0.152	N/A	N/A
rs12425705	chr12:50610321	*LIMA1* intron variant	4.01e-07	-0.152	0.085	1.12
rs11169322	chr12:50610976	*LIMA1* intron variant	4.01e-07	-0.152	0.085	1.12
rs8181679	chr12:50611020	*LIMA1* intron variant	1.95e-12	0.200	0.890	0.99
rs12424691	chr12:50611477	*LIMA1* intron variant	4.01e-07	-0.152	0.085	1.12
rs1362983	chr12:50614707	*LIMA1* intron variant	4.01e-07	-0.152	0.085	1.12
rs3812825	chr12:50616346	*LIMA1* intron variant	4.01e-07	-0.152	0.085	1.12
rs7314465	chr12:50623658	*RP3-405J10.3-001* non coding transcript exon variant	4.01e-07	-0.152	0.085	1.12
rs7136648	chr12:50624822	*LIMA1* intron variant	2.13e-11	0.196	0.772	0.98
rs10783342	chr12:50628466	*LIMA1* intron variant	1.95e-12	0.200	0.815	0.98
rs11169332	chr12:50629612	*LIMA1* intron variant	4.01e-07	-0.152	0.085	1.12
rs10747573	chr12:50633839	*LIMA1* intron variant	4.80e-13	0.188	N/A	N/A
rs11169335	chr12:50636364	*LIMA1* intron variant	4.01e-07	-0.152	0.091	1.12
rs12828340	chr12:50637295	*LIMA1* intron variant	4.01e-07	-0.152	0.085	1.12
rs7957659	chr12:50638810	*LIMA1* intron variant	1.95e-12	0.200	0.862	0.99
rs7953953	chr12:50647224	*LIMA1* intron variant	4.01e-07	-0.152	0.079	1.13
rs7486747	chr12:50650564	*LIMA1* intron variant	4.01e-07	-0.152	0.079	1.13
rs6580735	chr12:50665227	*LIMA1* intron variant	4.01e-07	-0.152	0.068	1.13
rs11169348	chr12:50665946	*LIMA1* intron variant	4.01e-07	-0.152	0.068	1.13
rs2111988	chr12:50668538	*LIMA1* intron variant	4.01e-07	-0.152	0.068	1.13
rs11169351	chr12:50672214	*LIMA1* intron variant	4.01e-07	-0.152	0.068	1.13
rs7967954	chr12:50673484	*LIMA1* intron variant	1.95e-12	0.200	0.768	0.98
rs10876014	chr12:50674753	*LIMA1* intron variant	4.01e-07	-0.152	0.068	1.13
rs10876015	chr12:50677506	Intergenic	4.01e-07	-0.152	0.074	1.13

rs200533278	chr12:50678972	Intergenic	4.01e-07	-0.152	N/A	N/A
rs6580736	chr12:50679418	Intergenic	3.66e-06	-0.140	0.121	1.11
rs10876017	chr12:50681539	Intergenic	4.01e-07	-0.152	0.068	1.13
rs11838347	chr12:50687160	Intergenic	4.01e-07	-0.152	0.068	1.13
rs11169370	chr12:50705872	Intergenic	9.03e-07	-0.147	0.068	1.13
rs35663729	chr12:50708870	Intergenic	4.01e-07	-0.152	0.068	1.13
rs7310541	chr12:50725965	*FAM186A* intron variant	4.01e-07	-0.152	0.074	1.13
rs6580741	chr12:50727706	*FAM186A* missense variant	4.01e-07	-0.152	0.080	1.13
rs7134595	chr12:50730458	*FAM186A* intron variant	4.01e-07	-0.152	0.073	1.13
rs4768900	chr12:50734199	*FAM186A* intron variant	4.01e-07	-0.152	0.068	1.13
rs4768951	chr12:50739008	*FAM186A* intron variant	3.66e-06	-0.140	0.112	1.11
rs7295847	chr12:50743913	*FAM186A* intron variant	4.01e-07	-0.152	0.068	1.13
rs7296291	chr12:50744119	*FAM186A* missense variant	4.01e-07	-0.152	0.068	1.13
rs7312252	chr12:50744171	*FAM186A* synonymous variant	4.01e-07	-0.152	0.068	1.13
rs10506292	chr12:50744753	*FAM186A* synonymous variant	4.01e-07	-0.152	0.068	1.13
rs4421818	chr12:50749294	*FAM186A* synonymous variant	4.01e-07	-0.152	0.068	1.13
rs12303082	chr12:50754563	*FAM186A* missense variant	4.01e-07	-0.152	0.068	1.13
rs11833608	chr12:50757628	*FAM186A* intron variant	4.01e-07	-0.152	0.068	1.13
rs10876027	chr12:50763484	*FAM186A* intron variant	3.66e-06	-0.140	0.124	1.11
rs12582180	chr12:50767285	*FAM186A* intron variant	3.66e-06	-0.140	0.114	1.11
rs7136702	chr12:50880216	Intergenic	1.02e-05	-0.129	0.053	1.14
rs9788075	chr12:51019171	*DIP2B* intron variant	3.79e-07	-0.146	0.111	1.11
rs10876074	chr12:51031817	*DIP2B* intron variant	3.79e-07	-0.146	0.111	1.11
rs1316607	chr12:51042890	*DIP2B* intron variant	3.79e-07	-0.146	0.106	1.11
rs4768903	chr12:51045449	*DIP2B* intron variant	2.81e-06	-0.136	0.126	0.90
rs7309964	chr12:51064064	*DIP2B* intron variant	4.02e-06	-0.142	0.031	1.16
rs11169520	chr12:51073523	*DIP2B* non-coding transcript exon variant	4.02e-06	-0.142	0.029	1.17
rs12427378	chr12:51074199	*DIP2B* intron variant	3.79e-07	-0.146	0.121	1.11
rs2090852	chr12:51086931	*DIP2B* intron variant	3.79e-07	-0.146	0.104	1.12

rs2139930	chr12:51089287	*DIP2B* intron variant	3.79e-07	-0.146	0.113	1.11
rs11169524	chr12:51089734	*DIP2B* synonymous variant	3.79e-07	-0.146	0.082	1.12
rs3742062	chr12:51128832	*DIP2B* intron variant	7.72e-06	-0.134	0.080	1.13
rs2280503	chr12:51138687	*DIP2B* 3' UTR variant	7.72e-06	-0.134	0.074	1.13
rs61926301	chr12:51157863	*ATF1* 5' UTR variant	4.05e-06	-0.135	0.100	1.11
rs12372718	chr12:51171090	*ATF1* intron variant	4.05e-06	-0.135	0.112	1.11
rs10783387	chr12:51180143	*ATF1* intron variant	4.05e-06	-0.135	0.112	1.11
rs7133974	chr12:51184577	*ATF1* intron variant	4.05e-06	-0.135	0.112	1.11
rs1129406	chr12:51203371	*ATF1* intron variant	4.05e-06	-0.135	0.116	1.11
rs4986838	chr12:51203376	*ATF1* synonymous variant	4.05e-06	-0.135	0.116	1.11
rs11169567	chr12:51204938	*ATF1* intron variant	4.05e-06	-0.135	0.116	1.11
rs7306677	chr12:51205763	*ATF1* intron variant	4.05e-06	-0.135	0.116	1.11
rs11169571	chr12:51213765	*ATF1* 3' UTR variant	4.05e-06	-0.135	0.116	1.11
rs10876098	chr12:51220373	Intergenic	4.05e-06	-0.135	0.074	1.13

Table 6.7 Variants that are more significantly associated with PBMC *CERS5* expression than the tagging SNP rs7136702 are shown in the table, listed in order of their chromosomal positions. The tagging SNP is included and highlighted in brown for reference; SNPs within the eQTL peak are in blue, whereas SNPs better associated with CRC risk are in yellow.

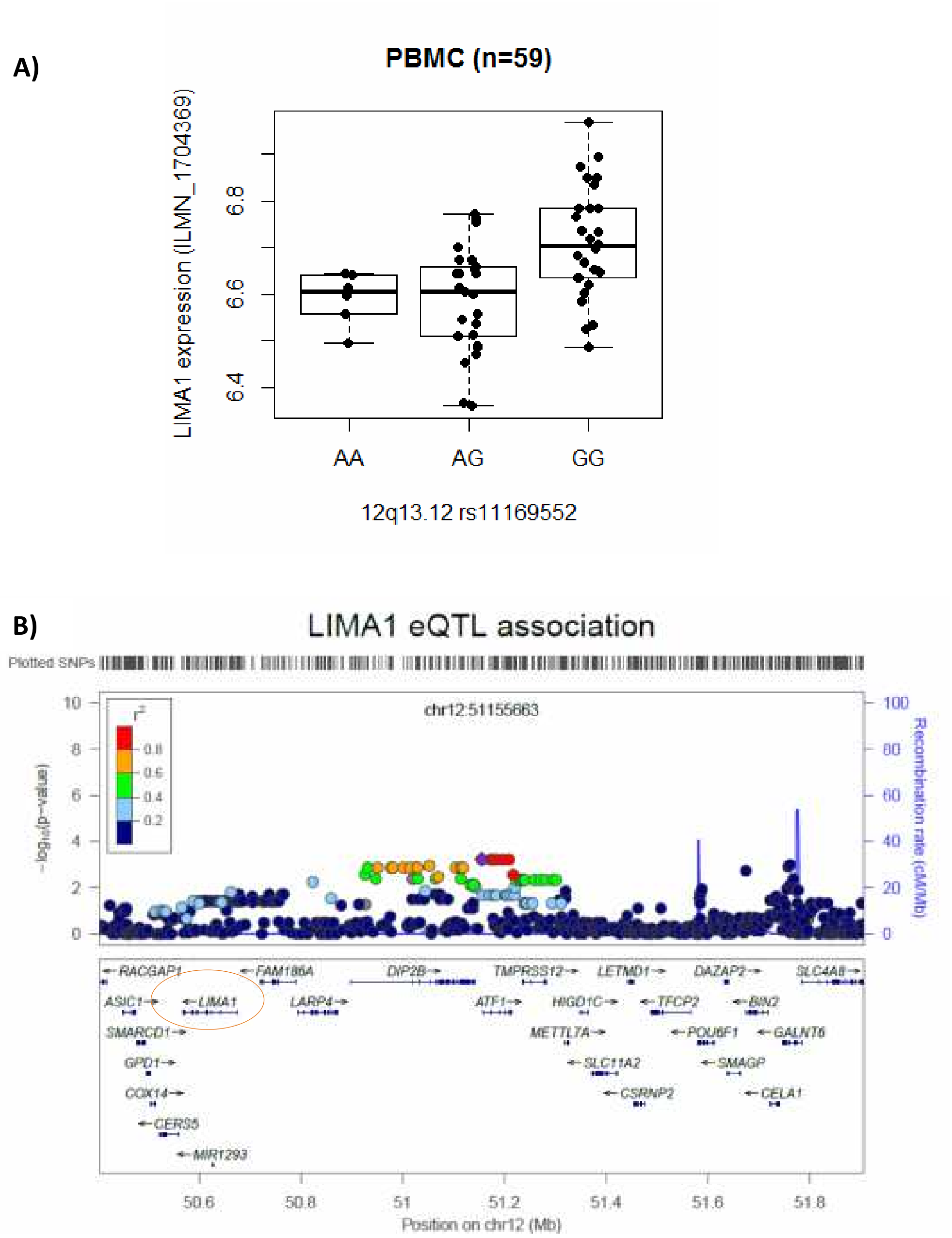


Figure 6.8 A) Boxplot showing the rs11169552 genotype association with *LIMA1* expression in PBMC. B) Fine association mapping for *LIMA1* expression detailing a 800kb region in LD with rs11169552 (purple).

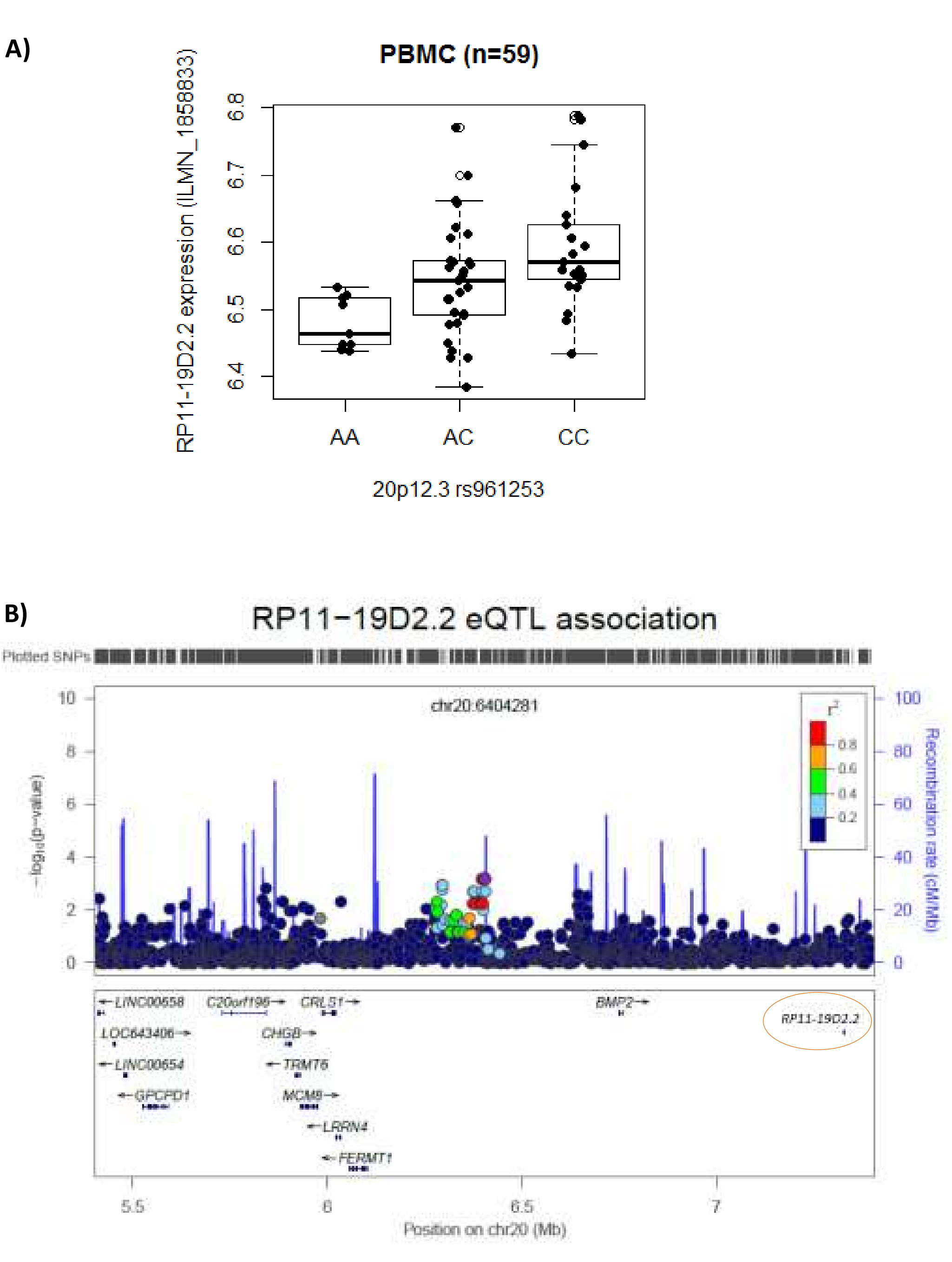


Figure 6.9 A) Boxplot showing the rs961253 genotype association with *RP11-19D2.2* expression in PBMC. B) Manhattan plot demonstrating *RP11-19D2.2* location in relation to the eQTL locus.

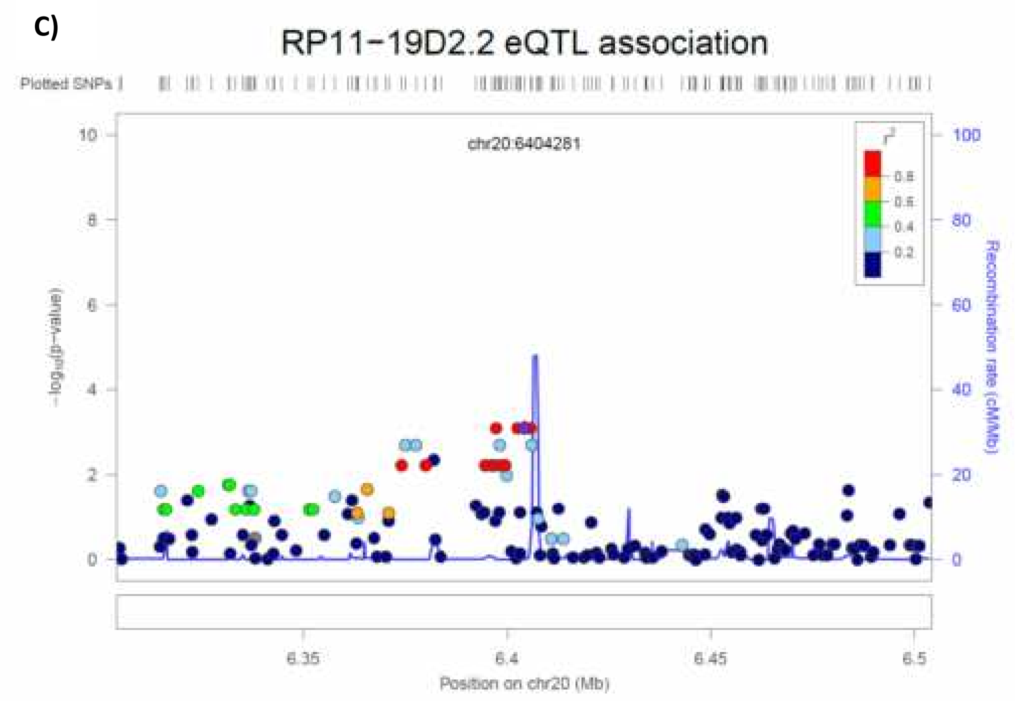


Figure 6.9 C) Fine association mapping for *RP11-19D2.2* expression detailing a 100Kb region in LD with rs961253 SNP (purple).

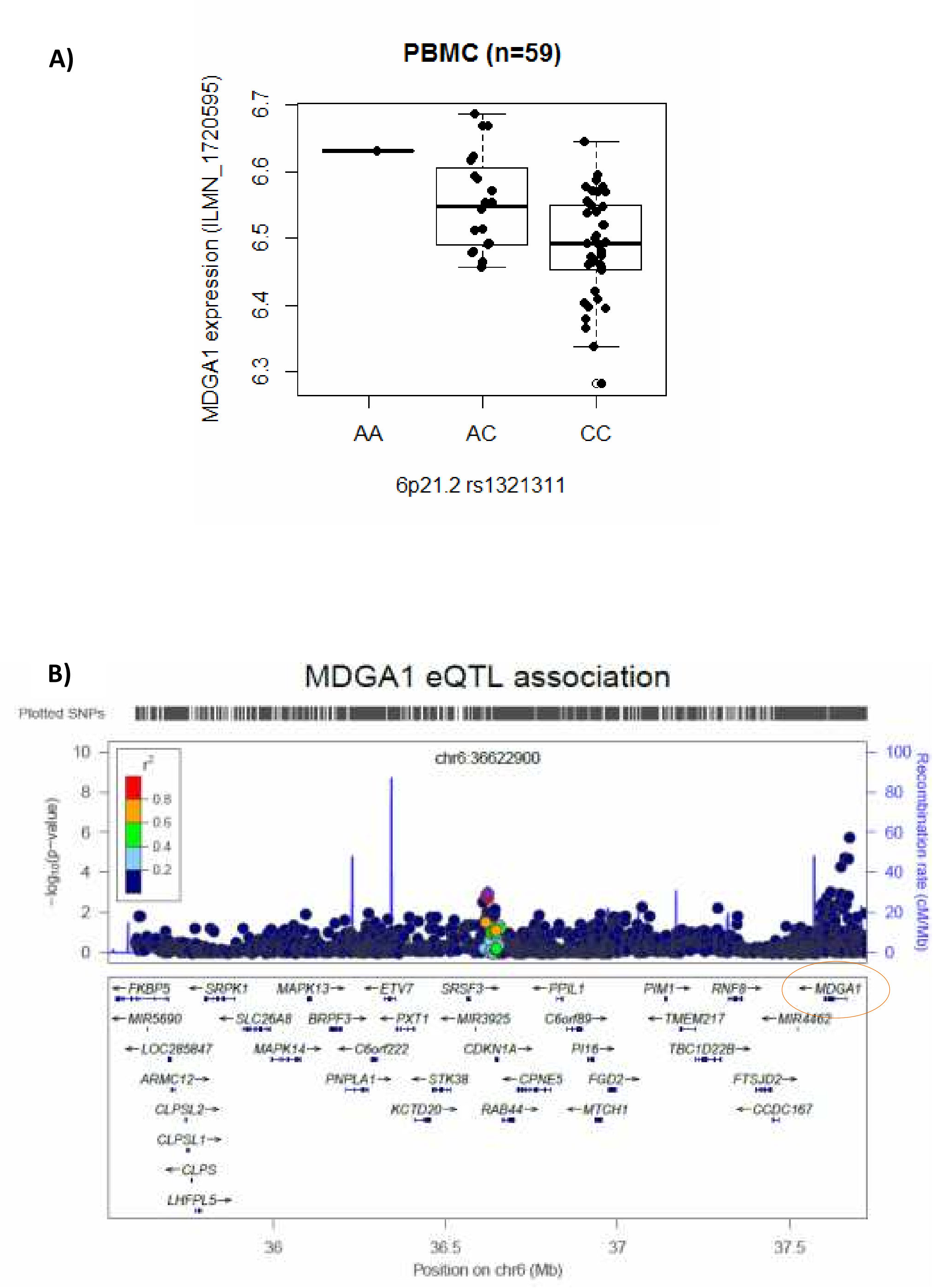


Figure 6.10 A) Boxplot showing the rs1321311 genotype association with *MDGA1* expression in PBMC. B) Manhattan plot demonstrating *MDGA1* location in relation to the eQTL locus.

C)

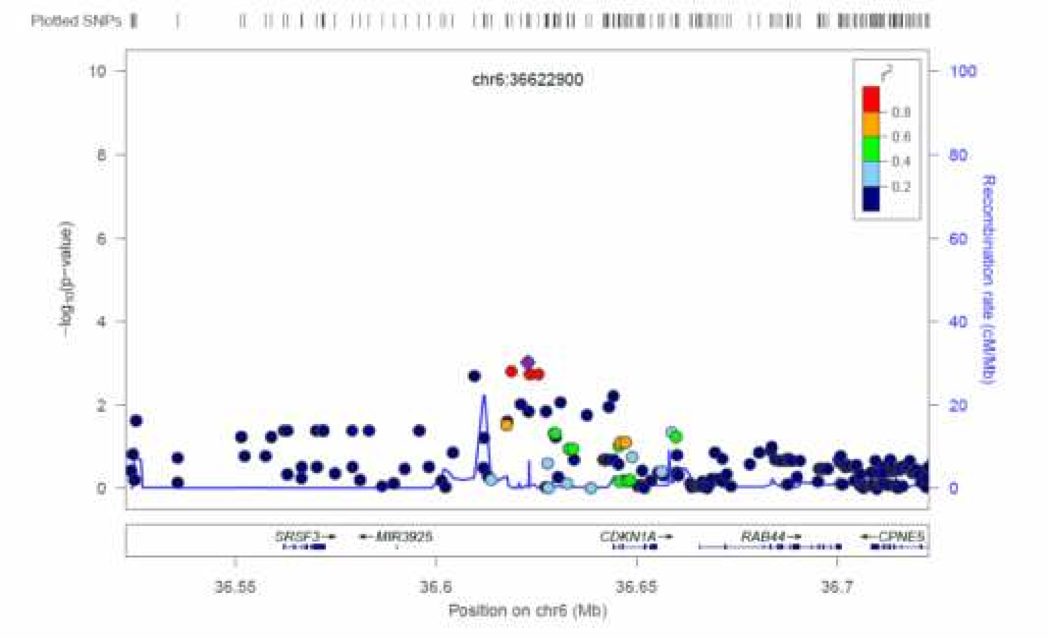


Figure 6.10 C) Fine association mapping for *MDGA1* expression detailing a 50Kb region in LD with rs1321311 SNP (purple).

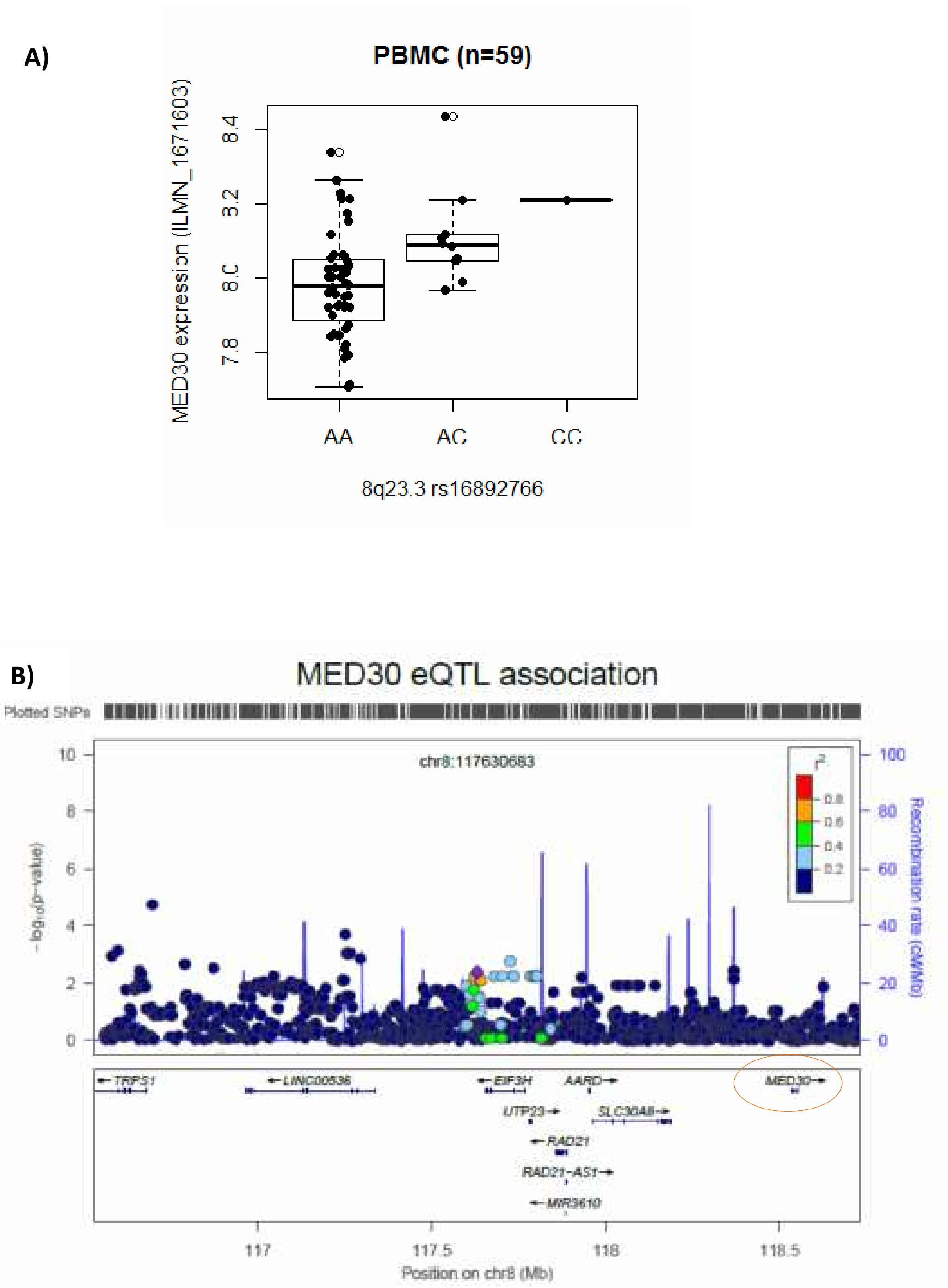


Figure 6.11 A) Boxplot showing the rs16892766 genotype association with *MED30* expression in PBMC. B) Manhattan plot demonstrating *MED30* location in relation to the eQTL locus.

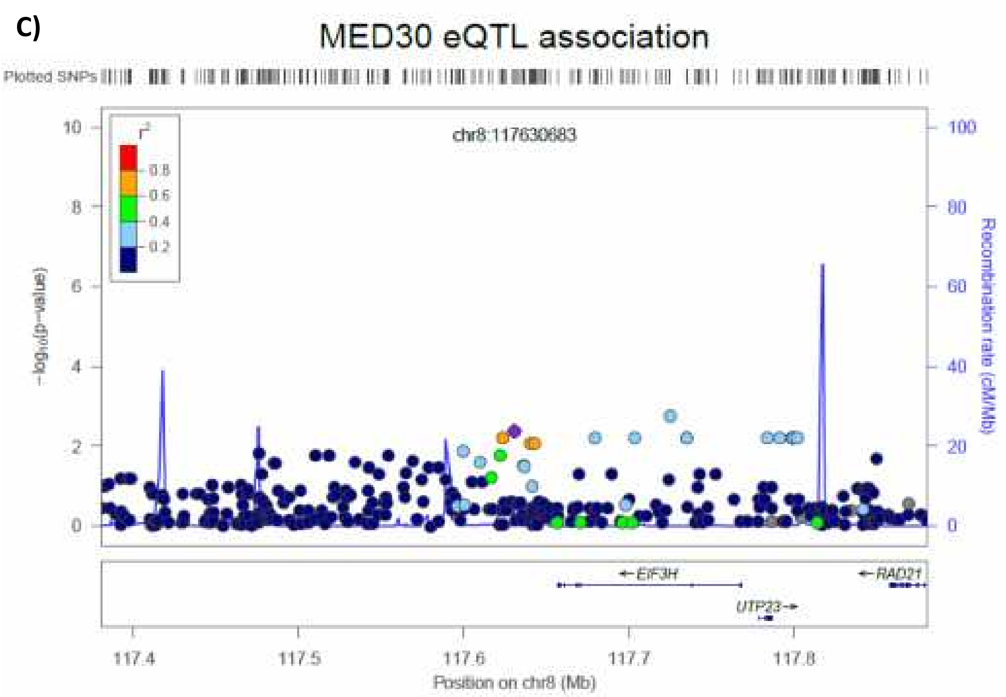


Figure 6.11. C) Fine association mapping for *MED30* expression detailing a 200Kb region that is in LD with rs16892766 (purple).

			eQTL		Case-control	
SNP	**SNP position**	**Predicted function**	***p*-value**	**beta**	***p*-value**	**OR**
rs16892766	chr8:117630683	Intergenic	4.15E-03	0.122	0.074	1.22
rs7825662	chr8:117725175	EIF3H intron variant	1.68E-03	0.127	0.755	1.03

Table 6.8 Variant that is more significantly associated with PBMC *MED30* expression than the tagging SNP rs16892766 are shown in the table. The tagging SNP is highlighted for reference. *p*-values and effect sizes for their eQTL association and CRC risk (case-control comparison) are presented.

6.3.4 Validation of the rs5934683-*SHROOM2* expression association

Same-sample technical validation of the rs5934683 eQTL was performed with qRT-PCR, measuring the same mRNA samples (n=115) used for whole-genome expression profiling. There was a strong correlation between *SHROOM2* expression on the Illumina HT12 microarray and expression measured by qRT-PCR (*p*-val<2.2e-16, spearman rho=0.66) (Figure 6.12). There was also a highly significant association (*p*-value=2.59e-07) between rs5934683 and *SHROOM2* expression measured by qRT-PCR (Figure 6.13A), validating the eQTL association seen with the Illumina HT12 microarray. The SNP accounted for 22% of the variability in *SHROOM2* expression, which is indicated by the coefficient of determination R^2 in the linear model. The risk allele T is associated with lower expression of *SHROOM2*, with a fold difference between the homozygotes/hemizygotes for the risk allele and the protective allele of 2.75 (95% CI, 1.96 – 4.36). Interestingly, in this linear regression model, gender appeared to have a significant influence on *SHROOM2* expression (*p*=0.003). This was independent of rs5934683 genotype, as there was no significant statistical interaction between the SNP genotype and gender (*p*=0.55). This gender-specific difference in *SHROOM2* expression will be discussed further in the next section.

On the other hand, qRT-PCR of the same PBMC samples (n=59) confirmed that *SHROOM2* was very lowly expressed and there was no detectable association (*p*=0.37) between rs5934683 genotype and *SHROOM2* expression (Figure 6.13B). There was also no differential expression between genders in PBMC *SHROOM2* expression (*p*=0.48).

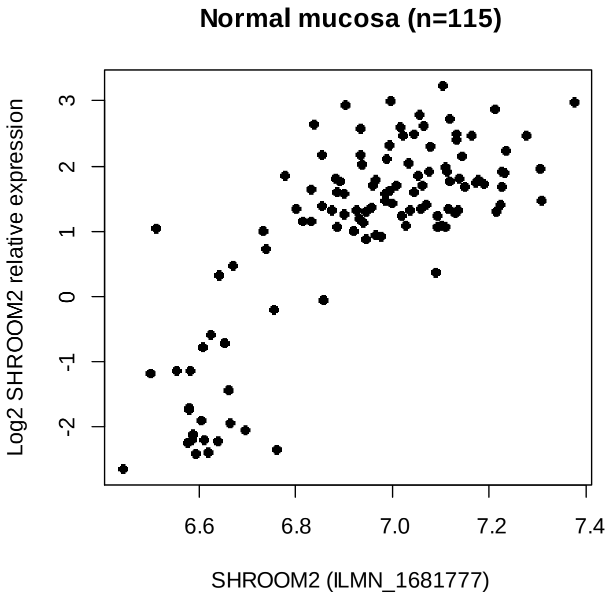


Figure 6.12 Significant correlation was observed between the expression of *SHROOM2* as quantified by Illumina HT12 microarray and qRT-PCR (*p*<2.2e-16, spearman rho=0.66).

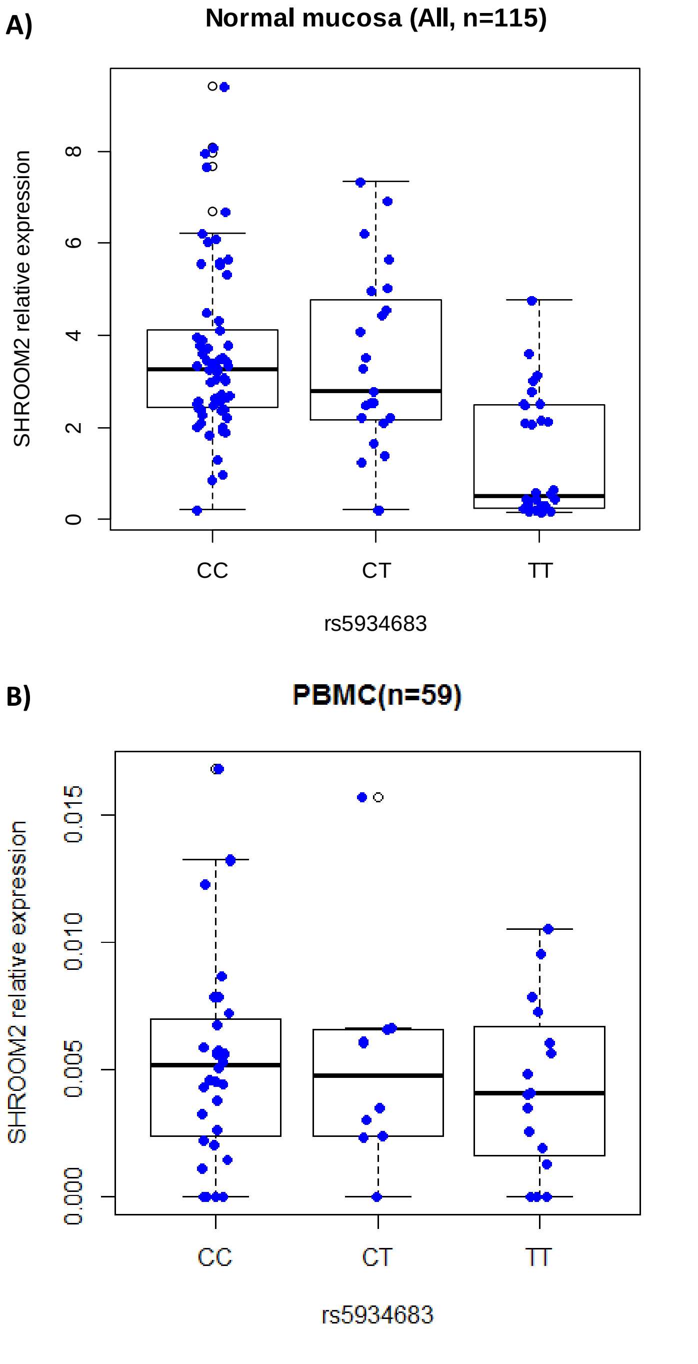


Figure 6.13 **A)** The association between rs5934683 and *SHROOM2* expression in the normal mucosa, as quantified by qRT-PCR. Linear regression adjusted for age, gender and anatomical site (*p*=2.59e-07; R^2 for rs5934683 = 0.215). **B)** *SHROOM2* relative expression in PBMC as quantified by qRT-PCR (normalised to *GAPDH*). Linear regression adjusted for age and gender (*p*=0.37).

To seek further confidence in the rs5934683-*SHROOM2* eQTL association, further replication was introduced at the level of normal mucosa sampling and RNA extraction. In a subset of 37 patients, 2 further RNA extracts were prepared from the normal mucosa harvested at the same time as the first extract but with spatial variation within the same colonic site. This spatial replication is thought to be of particular importance in the females, as heterozygous X-linked polymorphisms are functionally mosaic and the progeny of a single X-inactivation pattern are arranged together as large patches in the colon. The association between rs5934683 and *SHROOM2* was significant (*p*=3.77e-04) (Figure 6.14), and this remained true when the genders were analysed separately (males, *p*=1.10e-03; females, *p*=2.99e-03) (Figure 6.15).

In summary, the eQTL association between rs5934683 and *SHROOM2* expression in the normal mucosa was technically validated with qRT-PCR, and successfully replicated by multiple sampling in a subset of 37 patients. This association was tissue-specific and was not seen in the qRT-PCR validation of PBMC *SHROOM2* expression.

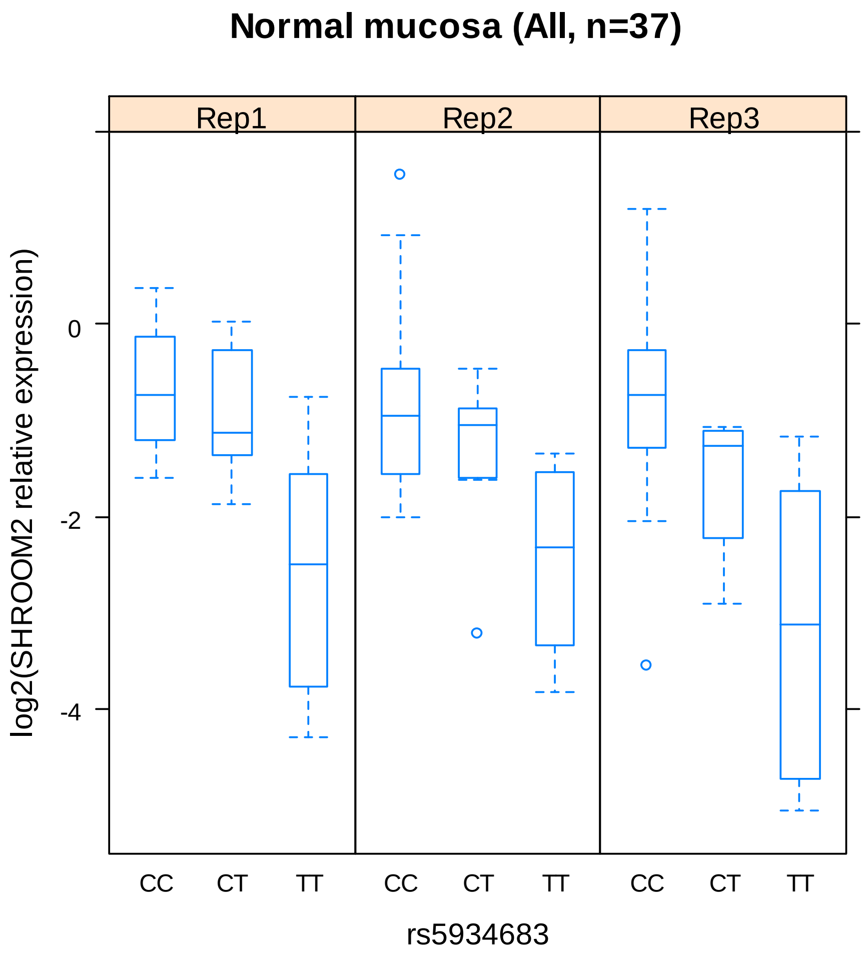


Figure 6.14 Biological replication of the association between *SHROOM2* and rs5934683 genotype. *SHROOM2* expression was measured with qRT-PCR in 3 different extracts of normal mucosa tissue taken at the same time (nested one-way ANOVA, *p*=3.77e-04).

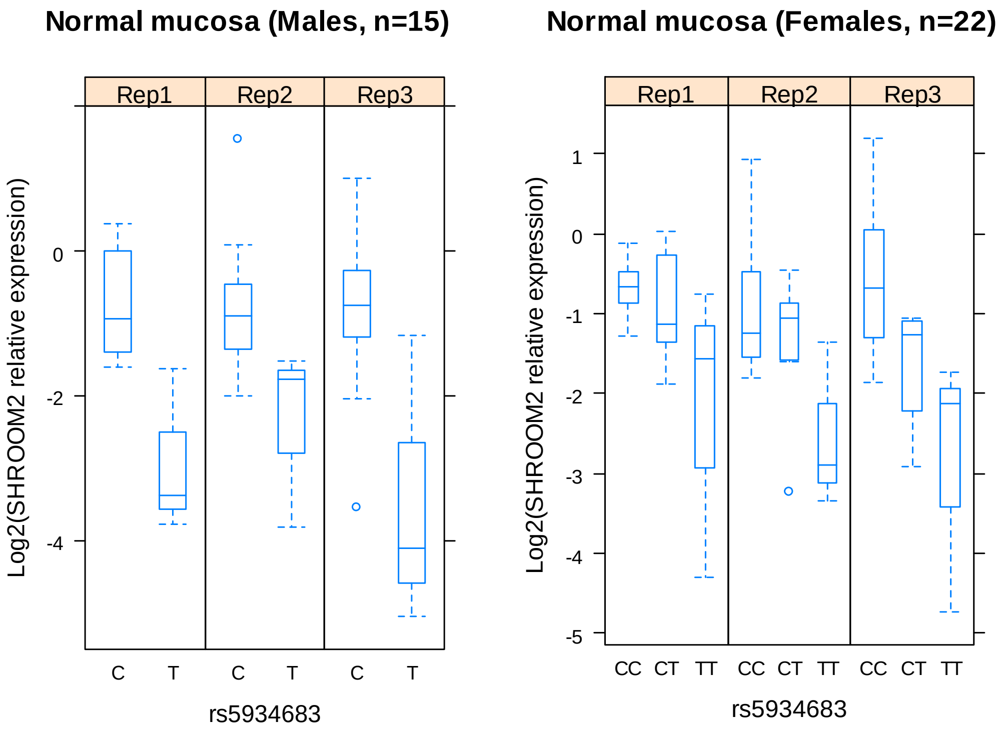


Figure 6.15 Biological replication of the association between *SHROOM2* expression and rs5934683 genotype, genders analysed separately with nested one-way ANOVA (males, *p*=1.10e-03; females, *p*=2.99e-03).

6.3.5 Gender-specific differences in *SHROOM2* expression

In the linear regression modelling for *SHROOM2* expression as quantified by qRT-PCR, it was observed that gender was significantly associated with *SHROOM2* expression (Figure 6.16). Overall, expression appears to be higher in females (*p*=0.003), with a mean fold increase of 1.46 compared to males (95% CI, 1.15-1.89). This differential expression was not observed in the Illumina microarray expression data. This discrepancy could be a result of the detection of differing transcript isoforms, or that of the limitations known to accompany microarray experiments. It has been recognised that microarrays tend to have lower sensitivities for certain genes (Chuaqui *et al*, 2002), with a significant decrease in overall accuracy of differential expression detection at low expression level and relatively poor sensitivity in detecting fold changes of less than 2 (Wang *et al*, 2006). It is plausible that the lack of association on the microarray is a false negative, as the fold change is small and *SHROOM2* is relatively lowly expressed.

In view of the gender difference in *SHROOM2* expression, I analysed the rs5934683 eQTL association separately in males and females (Figure 6.17). Although this association was still significant in both genders, it was considerably weaker in the females (males, *p*=3.23e-06; females, *p*=0.027). Not only was the association less significant in females, the variability in expression that was accounted for by the SNP was only 9% in females compared to 31% in males.

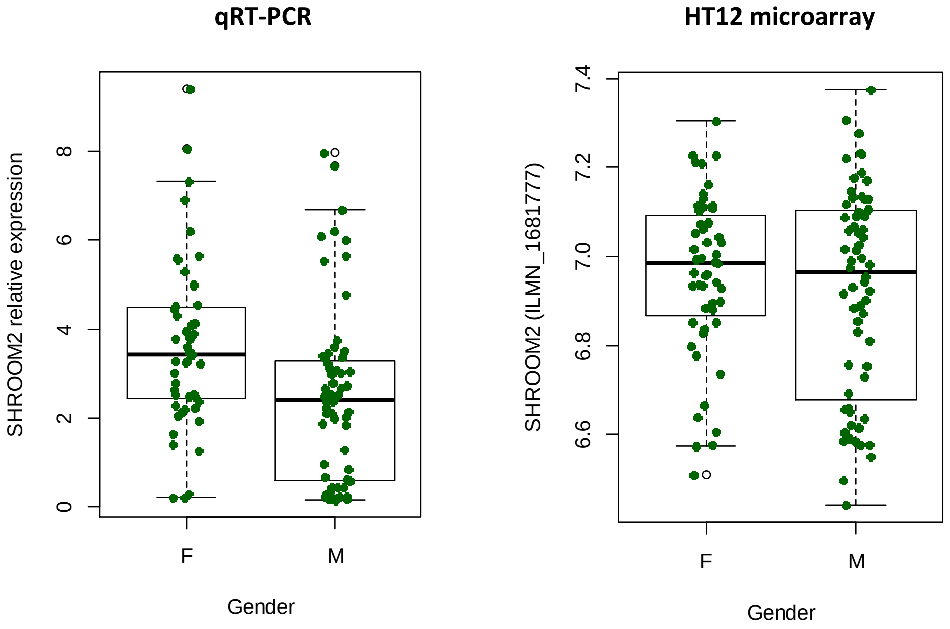


Figure 6.16 Comparison of *SHROOM2* expression between genders. There was significant differential expression in *SHROOM2* quantified by qRT-PCR. Linear regression adjusted for age, anatomical site and rs5934683 genotype (*p*=0.003, R^2 for gender=0.080). *SHROOM2* expression was higher in females with mean fold change=1.46, 95% CI [1.15-1.89]). This relationship was not seen in the *SHROOM2* expression data from the Illumina HT12 microarray (*p*=0.515).

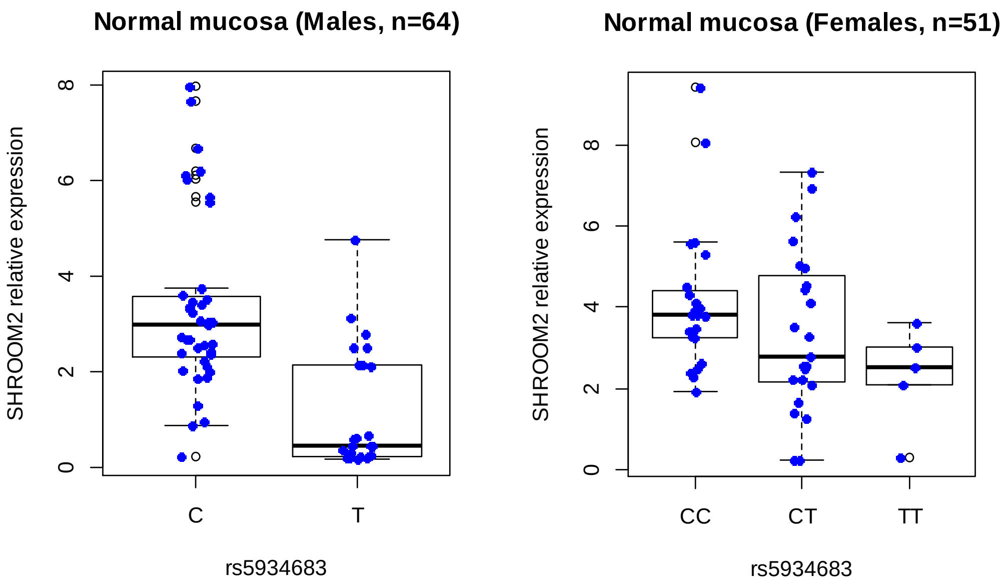


Figure 6.17 qRT-PCR validation of the association between rs5934683 and *SHROOM2* microarray expression, analysed separately by gender. Linear regression, adjusted for age and anatomical site (males, *p*=3.23e-06, R^2 for rs5934683=0.313; females, *p*=0.027, R^2 for rs5934683=0.086)

6.4 Discussion

By using whole-genome expression profiling and eQTL analysis of fresh normal mucosa samples and PBMC, I have demonstrated that a number of CRC risk variants are eQTLs that are associated with the expression levels of local *cis*-genes. For each of these loci, fine association mapping to the target gene expression levels and colorectal cancer risk was performed. This approach has revealed nearby SNPs in LD with the tagging SNP that are more highly associated with expression and risk, which makes them more likely to be the causal SNPs. By association, the target genes of these risk loci with eQTL activity are candidate susceptibility genes that may relevant to the predisposition and development of CRC.

The 11q23.1 locus demonstrated the strongest eQTL association, influencing the expression of two neighbouring genes, *COLCA2* and *COLCA1*, in the normal mucosa. The tagging SNP rs3802842 was one of the early GWAS discoveries (Tenesa *et al*, 2008) with an OR of 1.11, and has been replicated in subsequent studies and meta-analyses (Pittman *et al*, 2008; von Holst *et al*, 2010; Zou *et al*, 2012). The eQTL effects of this locus has recently been reported (Biancolella *et al*, 2014; Peltekova *et al*, 2014, Closa *et al*, 2014), providing independent validation of the data presented here. The rs7130173 SNP has been proposed by these studies to be the causal variant, as it explained the highest proportion of variance of the gene expression. In agreement with these reports, my findings showed rs7130173 as the best eQTL variant, together with a perfect proxy rs3087967. However, they did not perform better than the tagging SNP in the case-control analysis; another candidate variant rs112138001 was more associated with risk. Further functional studies will be required to elucidate the causal variant(s), as well as the allele-specific regulatory mechanism that is influenced by the polymorphic variants in this region. The eQTL target genes, *COLCA1* and *COLCA2* have not been characterised in depth, and have only recently studied in relation to its association with this CRC risk locus (Peltekova *et al*, 2014). The authors of this study showed via immunohistochemistry that *COLCA1* is largely expressed in the lamina propria, but not in normal epithelial cells or epithelium-derived neoplastic cells, whereas *COLCA2* is expressed in both the epithelium and the lamina propria. Based on the localisation of these genes in various mucosal immune cells of the colon, they proposed an immuno-regulatory role of these genes in the predisposition to CRC. However, when tested in our hands, the antibody used in this study did not appear to exhibit the level of specificity required for localisation via immunostaining, and that there is evidence to suggest that these genes are more likely to be long non-coding RNAs instead of protein-coding (Smillie, pers. comm.)

The Xp22.2 risk locus is the first X-linked locus to be associated with colorectal cancer (Dunlop *et al*, 2012). Here we suggested that it was a very strong colonic mucosa-specific eQTL with association to neighbouring gene *SHROOM2*, and this has been validated with qRT-PCR and replicated by repeated sampling in a subset of patients. This eQTL association was recently reported in an independent study (Closa *et al*, 2014), in which the authors also found an association with *GPR143* which was not observed in my data. This discrepancy could be due to the different microarray platform used, where a different probe sequence may have detected alternative transcripts. Fine-mapping of this region revealed putative functional variants within intron1 of *SHROOM2*, and will require more in-depth functional characterisation.

SHROOM2 belongs to the SHROOM family of proteins, which are regulators of epithelial morphogenesis, characterized by their ability to bind F-actin and organise actomyosin networks (Dietz *et al*, 2006), which makes it an interesting candidate for further study given the contribution of the actin and microtubule cytoskeleton to the cell biology of cancer (as reviewed by Hall, 2009). The technical validation of the Xp22.2 eQTL effect highlighted a gender-specific differential expression of *SHROOM2* in the normal mucosa, where expression was overall higher in females. Although the eQTL association was significant in both genders, there was attenuation of this effect in females, suggesting involvement of gender-specific factors in the regulation of *SHROOM2*. This is of interest as it is known that gender significantly influences the clinical and pathological characteristics of CRC; not only does it impact on the age-standardised incidence and mortality rate (Regula *et al*, 2006; Brenner *et al*; 2007), it also influences where tumours arise within the colon (reviewed by Koo *et al*, 2010), with risk and outcomes more favourable for females than males. The Xp22 locus is rich in genes that normally escape X-inactivation (Carrel *et al*, 2005), hence it is possible incomplete X-activation of *SHROOM2* accounts for the higher expression in females and consequently contribute a protective effect against CRC. Escape mechanisms in X-inactivation leading to disease protection is not unprecedented; X-linked tumour suppressor genes (Zuo *et al*, 2007) and immunomodulatory genes (Anderson *et al*, 1999) have been identified, with skewing and leaky X-inactivation being hypothesised as mechanisms conferring a protective effect in females (Libert *et al*, 2010; Chaligné *et al*, 2014). Nevertheless, this is unlikely to be the sole mechanism of gender-specific expression, as other factors such as hormonally-driven regulatory elements are almost certainly involved too. Without further speculation at this juncture, it is suffice to say that the independent association of lower *SHROOM2* expression levels with two known CRC risk factors (gender = male; rs5934683 = T) makes it a compelling susceptibility gene. These observations suggest that *SHROOM2* may have a protective or tumour suppressive role, with lower expression levels increasing the risk of developing colorectal cancer.

The other gene that stands out from the local-eQTL analysis is *CERS5*. Not only is it the target gene of the strongest eQTL association in PBMC, its expression also appears to be weakly influenced by the rs7136702 risk variant in the normal mucosa, albeit not surviving the correction for multiple-testing. This variant appears to tag a very strong eQTL locus for *CERS5*, and fine-mapping of the wider LD block has also revealed several candidate variants that are more highly associated with both *CERS5* expression and CRC risk. *CERS5* (Ceramide synthase 5) is involved in the *de novo* synthesis of ceramide, a sphingolipid involved in cell death and proliferation. Ceramide synthases have been implicated in cancer and apoptosis, although the precise roles of distinct family members have not been fully understood. Interestingly, it has recently been demonstrated to be highly expressed in colorectal cancer tissue and is associated with poorer clinical outcomes (Fitzgerald *et al*, 2015).

The other candidate genes derived from the cis-eQTL analysis (*SPATS*2 and *UTP23* in the normal mucosa, *LIMA1*, *RP11-19D2.2*, *MDGA1* and *MED30* in PBMC) are perhaps less convincing candidates due to a weaker eQTL association with the respective risk loci, but may still be interesting candidates due to the known functions of their protein products. Of these, *LIMA1* (LIM domain and actin-binding protein 1) is the most interesting candidate. It was previously known as *EPLIN* (epithelial protein lost in neoplasm) when it was first identified to be a human epithelial cell protein that is down-regulated or lost in the majority of cancer cell lines and xenografts examined (Maul *et al*, 1999). It was later characterised as a cytoskeletal protein with actin-binding properties which links the cadherin-catenin complex to F-actin, stabilising the adherens junction in epithelial cells (Abe *et al*, 2008). However, the relevance of this gene to colorectal cancer is questionable, as the eQTL effect was only observed in PBMC and not in the normal mucosa. Although its role in non-epithelial cells is generally not well-studied, there is a recent report that *LIMA1* is targeted by *AP12-MALT1* (juxtaposition of apoptosis inhibitor 2 to MALT lymphoma translocation gene 1) (Nie *et al*, 2015), the most frequent recurrent chromosomal translocation present in lymphomas involving the mucosa-associated lymphoid tissue (MALT). The authors also showed that depletion of *LIMA1* in a B-cell derived cell line affected various cancer phenotypes such as growth and invasiveness, indicating a possible role of *LIMA1* dysregulation in B-cell lymphomagenesis. Although this suggestion of a possible role for *LIMA1* in intestinal immunity is intriguing, it is remains speculative to suggest a link with colorectal susceptibility, especially as it is unclear whether peripheral blood mononuclear cells are appropriate surrogates for mucosal immune cells.

MDGA1 (MAM Domain containing Glycosylphosphatidylinositol Anchor-1) is another interesting candidate with potential relevance to cancer biology, as there is evidence of its role in cell adhesion. It is a glycoprotein that is localised specifically into membrane lipid rafts, and contains structural features found in cell adhesion molecules. Cell line over-expression and knock-out studies suggests that *MDGA1* mediates cell-cell adhesion in a heterophilic manner by affecting adhesion to extracellular matrix proteins (Díaz-López *et al*, 2010). As with *LIMA1*, the *MDGA1* eQTL was detected in PBMCs and the function of the gene product is not well-studied in this tissue type.

Potentially more interesting in principle, is the risk variant association with *RP11-19D2.2*, an uncharacterised long intervening non-coding RNA. Non-coding RNAs exhibit cell-specific and developmental dynamic expression patterns capable of facilitating a wide repertoire of regulatory functions (Mercer *et al*, 2009); long non-coding RNAs (lncRNA) in particular can operate through a variety of mechanisms such as chromatin remodelling, transcriptional control, protein inhibition, post-transcriptional modifiers or decoy elements (reviewed by Cheetham *et al*, 2013), leading to alterations in expression profiles of various target genes involved in cell homeostasis and cancer progression. There is accumulating evidence linking the mis-expression of lncRNA to diverse cancers and implicating a role for them in cancer signalling pathways. Interestingly, there is already a report of a papillary thyroid cancer risk locus 14q13.3 influencing the transcription of a functional thyroid-specific lncRNA (*PTCSC3*) that has tumour suppressor properties (Jendrzejewski *et al*, 2012). The apparent gene desert region that includes the prostate cancer 8q24 locus have also been shown to produce a lncRNA that may be involved in prostate tumourigenesis (Chung *et al*, 2011). Using a genome-wide approach, another study has demonstrated tissue-dependent lncRNA cis-eQTLs, of which a proportion are also associated with complex traits and diseases (Kumar *et al,* 2013). It is likely that many more lncRNAs are transcribed from cancer risk loci, but it may require targeted interrogation as these low-abundance RNAs may not have been detected or annotated.

Invariably, there are caveats to consider at various stages of this study, in particular with regards to the study design and analytical methods. The small sample size may not be adequately powered to detection subtle *cis*-regulatory effects, particularly in the PBMC where only 59 samples were analysed. Indeed, this may be one of the reasons why the eQTL association between rs1321311 (6p21.2) and *CDKN1A* in lymphoblastoid cell lines and T-cells (Dunlop *et al*, 2012) was not replicated in my PBMC samples. PBMC are a heterogenous group of cells that consists of lymphocytes, monocytes and macrophages; this cellular heterogeneity could have also contributed to the variation ‘noise’, making the detection of eQTLs harder.

The concept that the quality of the study results is only as good as the quality of the samples resonated strongly, especially during the early stages of patient recruitment and tissue collection. Factors associated with the sampling procedure of the colonic tissue and blood can significantly affect downstream observations, and it is important to be aware of these at the start to reduce artefactual or confounding variability. For example, knowing that cancer field effects may potentially distort differential expression (Hawthorn *et al*, 2014), mucosa samples were harvested from the macroscopically normal resection margin furthest away from the tumour to reduce any field effects. The variability in tissue post-mortem and ischaemic time is another caveat, as this is dependent on several factors including the surgical procedure, the timing of the ligation of the vascular supply, and practical issues such as the availability of a pathologist. One may also argue whether these samples are truly baseline samples, given the inflammatory response that accompanies the trauma of abdominal surgery. This might be particularly relevant to the PBMC samples as a significant proportion of them were collected in the days following the operation, when reactive inflammatory responses are likely to peak. Future studies may benefit from pre-operative PBMC sampling. Other patient-dependent factors such as anaesthetic drugs, medications, diet and even stress levels can potentially affect gene expression, and are difficult to control for.

From the technical point of view, the use of gene expression microarrays for gene expression studies also comes with its own limitations. They are an excellent tool for initial target discovery, but the partial coverage, technical variability and the relatively limited dynamic range, places restraints on the technology with respect to sensitivity and specificity. Similarly, although the DNA arrays used in this study allows detailed coverage of common SNPs, they do not provide information on structural variation such as indel polymorphisms and copy number variants. With whole-genome and transcriptome sequencing technology becoming more accessible, there is huge potential and scope for these samples to be analysed with much more depth using an integrative approach, moving beyond eQTL cataloguing to high-resolution assessment of the transcriptome as a functional phenotype readout of genetic variation in the normal colonic mucosa. Recent RNA-seq studies in other tissue types have already shown that alternative isoform production (Lalonde *et al*,

2011) and variation in mRNA stability (Pai *et al*, 2012) are influenced by heritable genetic variation, and will be of definite interest in future studies.

In conclusion, the data presented here has provided evidence that a proportion of CRC genetic non-coding variants influence cancer predisposition, at least in part, by affecting the expression levels of candidate genes in two different tissue types – the colonic mucosa and peripheral blood mononuclear cells. Although there is evidence that there is some overlap of eQTL effects between the colonic mucosa and blood (i.e. rs7136702), this evidence is weak and most of the eQTLs observed in this data appears to be tissue-specific. Considering the caveats discussed, the relatively small sample size, and the cellular heterogeneity of the tissue substrate, the ability to detect eQTL effects is quite remarkable, but may not be entirely surprising, as other published studies suggest that eQTLs tend to explain a greater proportion of target gene expression variance than is typically seen for risk alleles and clinical traits. It should be noted that identification of an eQTL provides only indirect evidence of a link between genotype and gene transcription, experimental and molecular approaches are necessary for confirming its mechanistic relevance. Methods to elucidate the molecular mechanism of polymorphic cis regulation are not easily amenable to such high-throughput analyses, and will be the next key challenge in validating these eQTL findings. The two best risk loci showing local eQTL effects (11q23.3 and Xp22.2) were taken forward for functional studies. Collaboratively, the 11q23.3 locus was validated and interrogated by another PhD student in the group (Claire Smillie) whereas I focused on the characterisation of the Xp22.2 locus and *SHROOM2*, which will be described in the next two chapters.

Chapter 7

Identification of a novel indel polymorphism as the causal variant of the Xp22.2 colorectal cancer risk locus

7.1 Introduction

By using whole genome expression profiling of normal colorectal mucosa tissue from 115 patients, the X-linked CRC risk SNP rs5934683 has been shown to be a strong eQTL governing expression of the neighbouring gene *SHROOM2*. This was initially observed in 42 patients (Dunlop *et al*, 2012), and subsequently replicated when more samples were collected and added into the analysis. To verify this eQTL association, the next challenge would be to define the regulatory mechanism underlying this relationship and identify the causal variant. Delineating the functional impact of this common, low-penetrance variant will provide tangible understanding of the mechanism by which common genetic variation imparts disease risk, which can in turn inform rational development of preventative strategies.

7.2 Methodological overview

7.2.1 Targeted resequencing

Targeted re-sequencing was performed on blood genomic DNA extracted from a subset (n=50) of the 115 patients in the eQTL analysis (Chapter 6). Sanger sequencing was performed as described in 2.4.2 for total of 5kb upstream of the SHROOM2 TSS. [†]

7.2.2 Indel24 genotyping

The Indel24 site was amplified with PCR as described in 2.4.2 using the following primers[‡]. The product of the insertion allele is 185bp whereas the product of the deletion allele is 161bp.

Forward primer	CACCCACATCCCGCTGATTG
Reverse primer	CCTTACCAAGAGGCGAAGC

A FAM fluorescent tag was attached to the 5' end of the reverse primer to allow sizing and quantification of the amplified DNA fragments. The products were scanned with the ABI PRISM HT7900 (Life Technologies) and analysed with the GeneScan® Analysis Software.

7.2.3 Construction of Manhattan plot and LD plot

Manhattan plots of the eQTL fine-association mapping at Xp22.2 was generated with the web tool LocusZoom as described in 6.2.4. Linkage disequilibrium plot of the Xp22.2 locus was constructed using the Haploview programme from the Broad Institute website (Barrett *et al*, 2005).

[†] Sanger sequencing performed by Stuart Reid, technician, MRC Human Genetics Unit, IGMM.
[‡] PCR and genotyping performed by Stuart Reid, technician, MRC Human Genetics Unit, IGMM.

7.2.4 Luciferase reporter assays

To study the effects of the Xp22.2 polymorphic variants on transcriptional activity, gene elements containing the different alleles of the 3 candidate variants were purified and subcloned into firefly luciferase reporter expression vectors (See Figure 7.14). Cloning and generation of test plasmids were performed by Stuart Reid, CCGG technician and will not be described in detail here. In brief, genomic blood DNA from patients heterozygous for these variants were amplified using proof-reading Taq Polymerase (Promega) and cloned into the pGEMT Easy vector (Promega). After identification and verification by Sanger sequencing, these were cloned into the luciferase reporter vectors pGL2 or pGL4 (Promega). The test plasmids containing the different alelles were transfected using Lipofectamine™ 2000 (Life Technologies) in Opti-MEM® I Reduced Serum Medium (Life Technologies) into colorectal cancer and retinal epithelial cell lines[§] when they are at 80-90% confluence, according to the manufacturer’s protocol. Briefly, for each transfection sample, the test plasmid DNA (500ng for each well in a 6-well plate) and Lipofectamine™ 2000 was diluted separately in Opti-MEM® and allowed to stand at room temperature for 5 minutes. The two solutions were then mixed and incubated for 30 minutes at room temperature to allow complexes to form, prior to addition to wells containing cells in antibiotic-free medium. pCMV-β (generated by Laura Boyes and Susan Farrington, CCGG) was co-transfected as a control for transfection efficiency. Cells were incubated in antibiotic-free media at 37ºC in a humidified incubator (95% O2, 5% CO2), and harvested after 24-48 hours. Cell extracts were prepared using Cell Culture Lysis Reagent (Promega), followed by the Luciferase Assay (Promega) and β-galactosidase Enzyme Assay (Promega). Fluorescence from luciferase activity was measured with the DLRead Lumat LB9507 luminometer (EG&G Berthold), whereas β-galactosidase expression was quantified by the Multiskan MS microplate reader (Labsystems). The luciferase activity in each sample was normalised with β -galactosidase expression.

[§] Transfections in the retinal pigment cell lines were performed by Andrew McBride, PhD student

7.2.5 siRNA gene knockdown in cell lines

Cells were plated the day before and grown until 40-60% confluent prior to siRNA transfection. SiRNAs used are detailed in Table. Transfections were carried out with Lipofectamine™ 2000 (Life Technologies) in Opti-MEM® I Reduced Serum Medium (Life Technologies), according to the manufacturer’s protocol. In brief, siRNA and Lipofectamine™ 2000 are diluted in the appropriate amount of Opti-MEM® I separately and allowed to equilibrate for 5 minutes at room temperature. The two solutions were then mixed and incubated for 30 minutes at room temperature to allow complexes to form, prior to addition to wells containing cells in antibiotic-free medium. Cells were harvested for protein/RNA extraction or assayed after 48 hours of incubation at 37ºC in a humidified incubator (95% O2, 5% CO2). A dose-response is first performed to determine the lowest effective concentration of siRNA for each individual gene and cell line used (usually between 5-15 nM) before phenotype assays.

Gene	siRNA	Oligo ID	Sequence (5’ - 3’)
NF-YA	siRNA3	SASI_Hs01_00020331	CGAUGAAGAAGCAAUGACA
NF-YA	siRNA4	SASI_Hs01_00183592	CCAAUGGGACAUUGAUGAU
NF-YB	siRNA1	SASI_Hs02_00341025	GCAUGAAUGAUCAUGAAGA
NF-YB	siRNA2	SASI_Hs02_00341024	GAAGGAAAGACUGGUGAAA
Negative control	Scrambled	SIC001	

Table 7.1 The IDs and sequence of siRNAs (Sigma) used in this chapter.

7.2.6 Co-transfection with luciferase reporter plasmid and siRNA

For luciferase reporter plasmid and NFY siRNA co-transfections in MCF7 cell line, cells were plated the day before to achieve 40-60% confluence. siRNA transfections were performed as described in 7.2.5 for 24 hours prior to luciferase reporter plasmid DNA transfections as described in 7.2.4. Cells were harvested after incubation for a further 24 hours.

7.2.7 qRT-PCR

qRT-PCR was performed on cDNA synthesised from cell line or primary tissue RNA as described in Chapter 2. The Taqman® Gene Expression Assays used are listed in Table 7.2.

Gene	Assay ID	Probe sequence
NF-YA	Hs00953589_m1	TCCCCATATGCAGGATCCAAACCAA
NF-YB	Hs01105350_m1	CAACATCATATCAACAGATTTCTGG
SHROOM2	Hs01113636_m1	CTCCCGGTGATCGGCAATCACTGCT
CCNB1	Hs01030099_m1	CTGAGCCTATTTTGGTTGATACTGC
TBP	Hs00427620_m1	GCAGCTGCAAAATATTGTATCCACA
RPL30	Hs00265497_m1	TATCATTGATCCAGGTGACTCTGAC

Table 7.2 TaqMan® assay IDs of the genes of interest and reference genes quantified in this chapter.

7.2.8 Site-directed mutagenesis of the CCAAT box motifs

Site directed mutagenesis of the CCAAT box motifs within the insertion allele of the 83+Indel24 luciferase reporter vector was performed using QuikChange II Site-Directed Mutagenesis Kit (Agilent Technologies) as per manufacturer's instructions.[**] These vectors were then transfected as described in 7.2.4 into SW480 and MCF7 cell lines for luciferase reporter assays.

7.2.9 Case-control analysis

Case-control analysis of the rs5934683 tagging SNP and the two putative causative variants Indel24 and rs5934685 was performed in 687 cases and 873 controls from the SOCCS (Scottish Colorectal Cancer Susceptibility) study. The putative causal variant Indel24 was genotyped as described in 7.2.2,[††] whereas

[**] Site-directed mutagenesis performed by Stuart Reid, technician, MRC Human Genetics Unit, IGMM

[††] Indel genotyping performed by Stuart Reid, technician, MRC Human Genetics Unit, IGMM.

rs5934685 was imputed as described in 6.2.2.[‡‡] Subsequently, Indel24 was genotyped in a larger dataset derived from samples from across Scotland, England and Croatia (cases=8368, controls=6327; males=7846, females=6849).[§§] Similar to the analysis in 6.2.4, male hemizygotes were treated as homozygotes in this case-control analysis.[***]

7.2.10 Western blotting

Total protein and subcellular fractions were extracted as described in 2.5. Primary antibodies used are listed in Table 7.3.

Protein	Company	Catalogue no.	Type	Antibody dilution used
NF-YA (G-2)	Santa Cruz	#sc-17753	Mouse monoclonal	1:1000
NF-YB (FL-207)	Santa Cruz	#sc-13045	Rabbit polyclonal	1:1000
β-actin	Sigma	#A1978	Mouse monoclonal	1:5000

Table 7.3 Details of the antibodies and dilutions used in this chapter.

[‡‡] rs5934685 imputation and statistical analysis performed by Maria Timofeeva, Statistical Geneticist, MRC Human Genetics Unit, IGMM

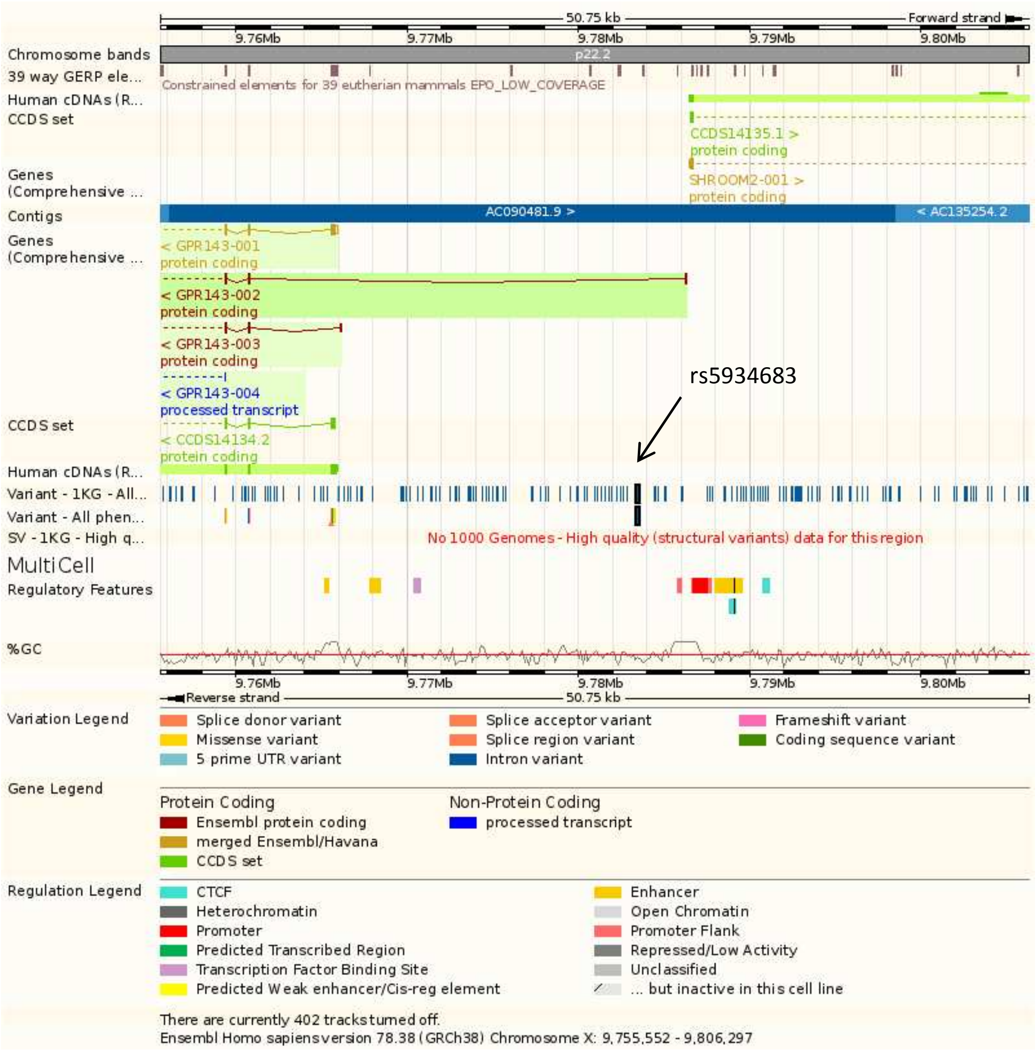
[§§] Indel24 genotyping performed by Stuart Reid, technician, MRC Human Genetics Unit, IGMM.

[***] Statistical analysis performed by Maria Timofeeva, Statistical Geneticist, MRC Human Genetics Unit, IGMM

7.3 Results

7.3.1 The genomic, epigenomic and regulatory landscape of rs5934683 from publicly available databases

The tagging SNP rs5934683 (chrX:9751474) resides within an intergenic region between the *GPR143* and *SHROOM2* genes at Xp22.2, which are divergently transcribed on opposite strands (Figure 7.1). The SNP is 3022bp from the 5' end of the *SHROOM2* canonical RefSeq gene structure, and is 17,469bp from the 5' end of the *GPR143* RefSeq gene structure. There is evidence of longer *GPR143* transcripts extending into the *SHROOM2* promoter, and rs5934683 SNP is within the first intron of this transcript (Ensemble transcript model ENST000000447366). The evidence for this transcript is weak though, with only a single EST supporting it from the HAVANA project (URL7.1).



There are currently 402 tracks turned off.
Ensembl Homo sapiens version 78.38 (GRCh38) Chromosome X: 9,755,552 - 9,806,297

Figure 7.1 The genomic context of rs5934683 from Ensembl (Genome assembly GRCh37). rs5934683 is highlighted in black in the variant track.

The eQTL activity associated with rs5934683 suggests that it may lie within or close to tissue specific regulatory elements. To look for evidence of this, regulatory data from the ENCODE project was first examined using the UCSC genome browser. Tracks examined include chemical modifications to histone proteins (H3K4Me1, H3K4Me3, H3K27Ac), DNase hypersensitivity, methylation and transcription factor binding. Bearing in mind that many regulatory elements appear to be tissue or cell type specific, it should be noted that the majority of the cell lines/cell types used to generate data for the ENCODE project are not of colonic origin. However, this information can still be inferential, as cis-eQTL datasets have been shown to overlap by more than 50% between cells as diverse as lymphoblastoid cells, hepatic cells and monocytes (Zeller *et al*, 2010).

The rs5934683 SNP appears to be encompassed within a DNase hypersensitive area of 350bp (chrX:9751266-9751615, Figure 2). Regulatory regions in general and promoters in particular, tend to be DNase-sensitive. However, the extent of the hypersensitivity is modest with a cluster score of 189/1000, and is only present in 5 cell types (H9ES, MCF-7, hepatocytes, myometrial cells, osteoblasts) out of the 125 tested. There are no relevant histone marks in the region; modifications to H3K4me1 only begin to become apparent ~1kb downstream and nearer to the *SHROOM2* promoter.

The closest transcription factor binding site (TFBS) is 972bp downstream and closer to the *SHROOM2* promoter (Figure 7.2). It is present in all three tested cell lines (GM12878, HeLa-S3 and K562), and binds to the transcription factor NFY-B. There is a moderately strong cluster score of 492 (out of 1000), with 3 common SNPs within the TFBS. The canonical DNA-binding motif for NFY-B has also been identified within the binding site by the Factorbook repository in-silico computational analysis (Wang *et al*, 2012; Wang *et al*; 2013). The ChIP-seq data from the previous ENCODE version (version 2) also shows a TFBS for NFY-A that overlaps with the NFY-B TFBS, which is not surprising given that the NFY transcription factor is a trimeric complex formed by the three subunits, NFY-A, NFY-B and NFY-C.

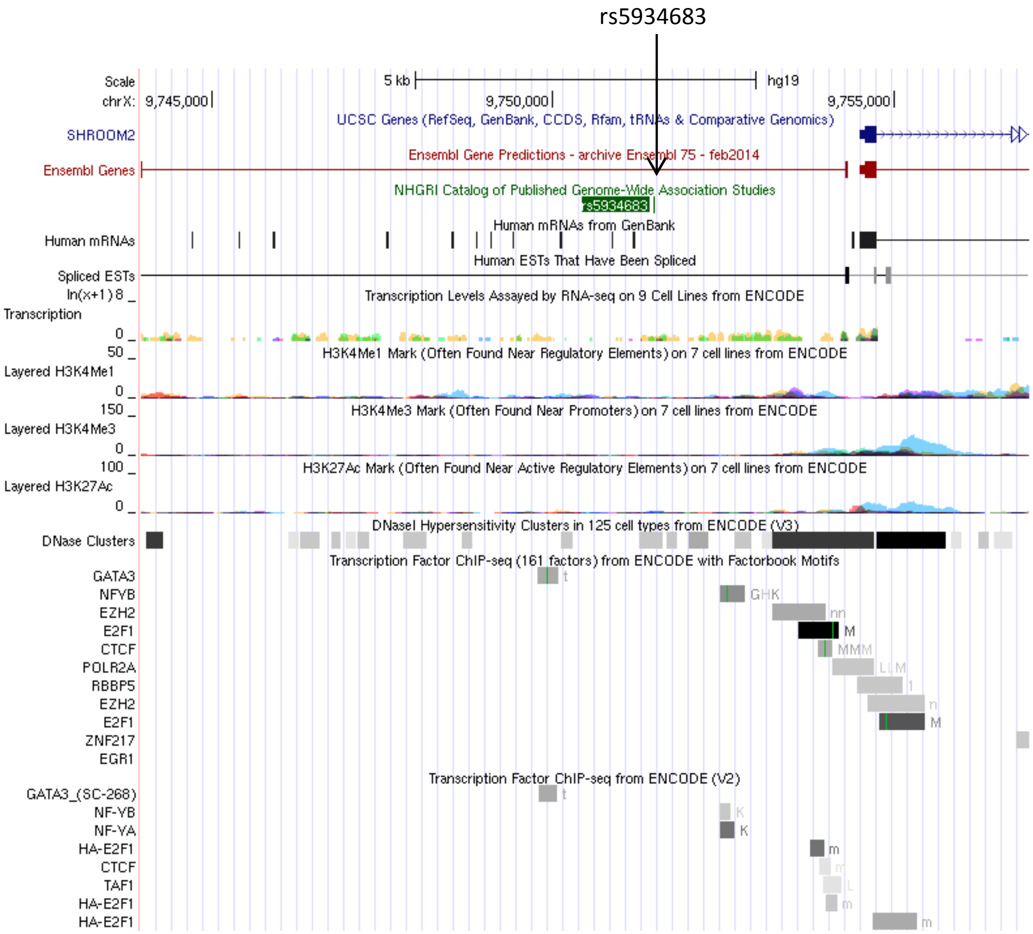

rs5934683
Scale
chrX:
9,745,000
5 kb
9,750,000
hg19
9,755,000
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
SHROOM2
Ensembl Gene Predictions - archive Ensembl 75 - feb2014
Ensembl Genes
NHGRI Catalog of Published Genome-Wide Association Studies
rs5934683
Human mRNAs from GenBank
Human mRNAs
Human ESTs That Have Been Spliced
Spliced ESTs
ln(x+1) 8
Transcription
0
Transcription Levels Assayed by RNA-seq on 9 Cell Lines from ENCODE
50
Layered H3K4Me1
0
H3K4Me1 Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
150
Layered H3K4Me3
0
H3K4Me3 Mark (Often Found Near Promoters) on 7 cell lines from ENCODE
100
Layered H3K27Ac
0
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
DNase Clusters
DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V3)
Transcription Factor ChIP-seq (161 factors) from ENCODE with Factorbook Motifs
GATA3
NFYB
EZH2
E2F1
CTCF
POLR2A
RBBP5
EZH2
E2F1
ZNF217
EGR1
t
GHK
nn
M
MMM
L M
1
n
M
Transcription Factor ChIP-seq from ENCODE (V2)
GATA3_(SC-268)
NF-YB
NF-YA
HA-E2F1
CTCF
TAF1
HA-E2F1
HA-E2F1
t
K
K
m
m
L
m
m

Figure 7.2 *SHROOM2* promoter region (UCSC browser, Hg19). Tracks displayed include transcription levels by RNA-seq, histone marks (H3K4Me1, H3K4Me3 and H3K27Ac), DNaseI hypersensitivity clusters and ChIP-seq transcription factor binding data.

A small, nucleosome depleted region (chrX:9751290-9751339) 135bp upstream of rs5934683 is covered by a probe from the Illumina Infinium Human Methylation 450 Bead Array platform (Figure 7.3). This probe appears to demonstrate differential DNA methylation, with data from GM12878, Hi-hESC, HeLa-S3, HepG2 and HUVEC indicating this region is fully methylated, and data from K562 indicating a lack of methylation. The same region shows evidence of methylation, apparently nucleated upon methylated CpG sites, in an independent sequencing based study of human frontal cortex (Maunakea *et al*, 2010). This region also shows evidence for association with the nuclear lamina: a chromatin state that is known to include regions with methylated CpG sites (Guelen *et al*, 2008).

The presence of rs5934683 within a DNase hypersensitive region suggests it lies within the distal promoter of SHROOM2; its presence near a TFBS and a differentially methylated promoter site is consistent with the eQTL activity of this SNP and suggests possible mechanisms underlying this activity. This allele-specific regulation may be driven by rs5934683 itself, or any polymorphism that it tags. As the linkage disequilibrium of this region is poorly defined (Figure 7.4), this chapter will focus on characterising the variation in the region and identifying the causal variant by using a combination of expression association analysis and functional in-vitro assays.

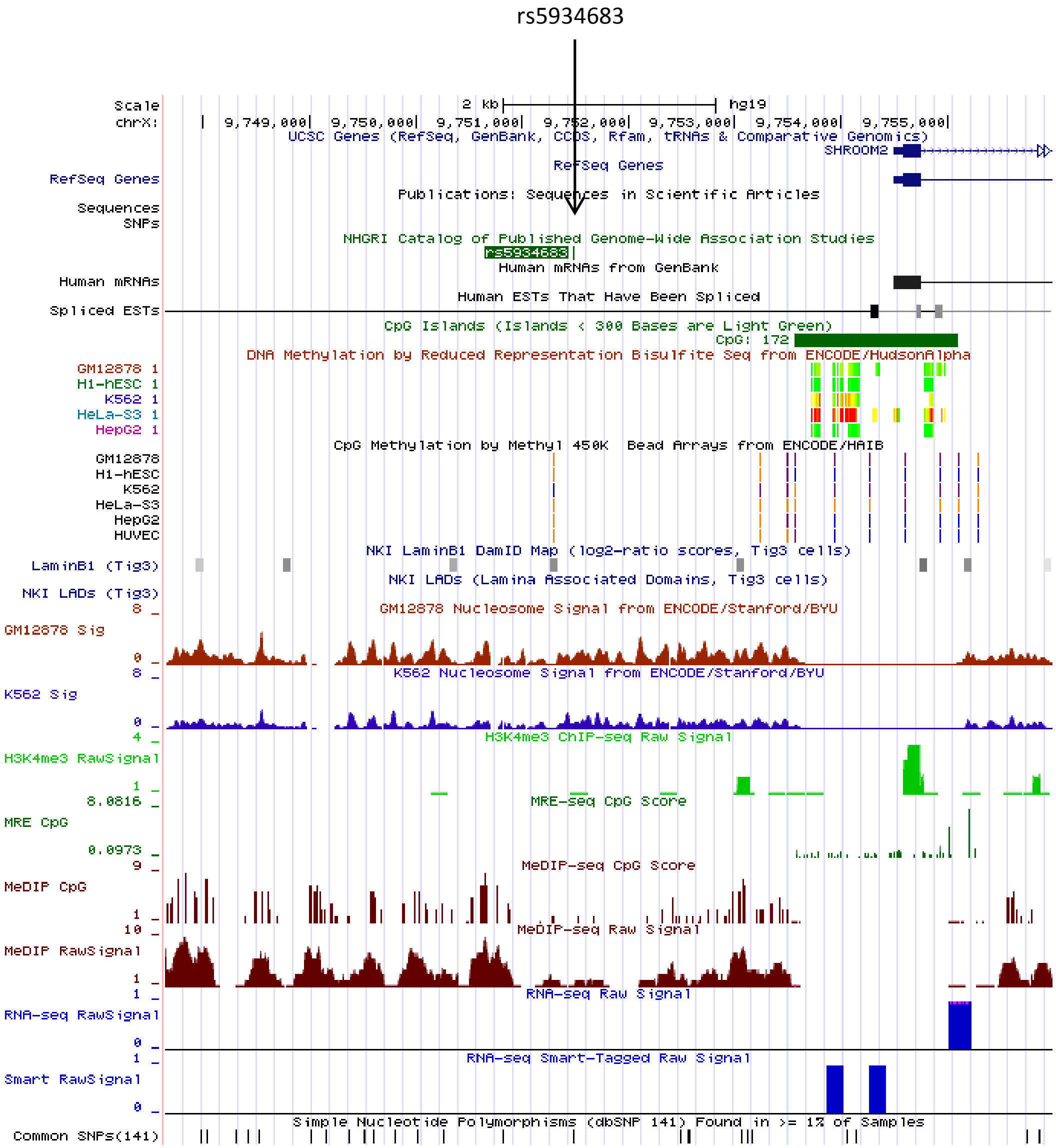

rs5934683
Scale
chrX:
2 kb
hg19
9,749,000| 9,750,000| 9,751,000| 9,752,000| 9,753,000| 9,754,000| 9,755,000|
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
SHROOM2
RefSeq Genes
RefSeq Genes
Publications: Sequences in Scientific Articles
Sequences
SNPs
NHGRI Catalog of Published Genome-Wide Association Studies
rs5934683
Human mRNAs from GenBank
Human mRNAs
Human ESTs That Have Been Spliced
Spliced ESTs
CpG Islands (Islands < 300 Bases are Light Green)
CpG: 172
DNA Methylation by Reduced Representation Bisulfite Seq from ENCODE/HudsonAlpha
GM12878 1
H1-hESC 1
K562 1
HeLa-S3 1
HepG2 1
CpG Methylation by Methyl 450K Bead Arrays from ENCODE/HAIB
GM12878
H1-hESC
K562
HeLa-S3
HepG2
HUVEC
NKI LaminB1 DamID Map (log2-ratio scores, Tig3 cells)
LaminB1 (Tig3)
NKI LADs (Lamina Associated Domains, Tig3 cells)
NKI LADs (Tig3)
GM12878 Nucleosome Signal from ENCODE/Stanford/BYU
GM12878 Sig
8
0
K562 Nucleosome Signal from ENCODE/Stanford/BYU
K562 Sig
8
0
H3K4me3 ChIP-seq Raw Signal
H3K4me3 RawSignal
4
1
MRE-seq CpG Score
MRE CpG
8.0816
0.0973
MeDIP-seq CpG Score
MeDIP CpG
9
1
MeDIP-seq Raw Signal
MeDIP RawSignal
10
1
RNA-seq Raw Signal
RNA-seq RawSignal
1
0
RNA-seq Smart-Tagged Raw Signal
Smart RawSignal
1
0
Simple Nucleotide Polymorphisms (dbSNP 141) Found in >= 1% of Samples
Common SNPs(141)

Figure 7.3 The *SHROOM2* promoter region (UCSC genome browser, Hg19). Tracks displayed include predicted CpG islands, methylation data from array and sequencing based assays, nucleosome occupancy, and lamin B1 association scores.

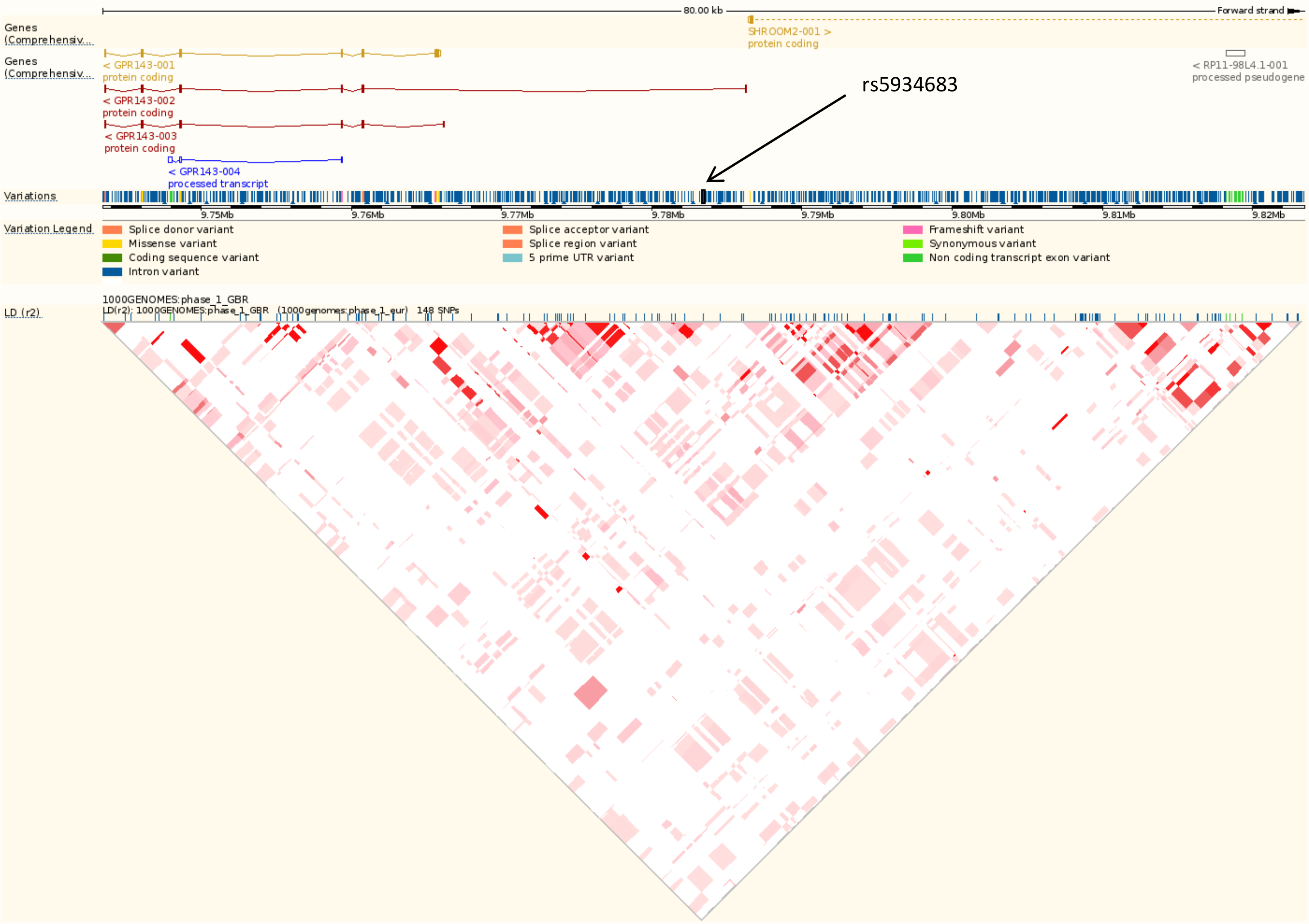


Figure 7.4. LD plot of the region surrounding rs5934683 derived from 1000GENOMES:phase_1_EUR (Ensembl, Genome assembly GRCh37). LD values (r^2) between any two variants are graphically displayed using inverted coloured triangles varying from white (low LD) to red (high LD).

7.3.2 Identification of putative causal variants by targeted resequencing and fine-mapping of the Xp22.2 locus

Targeted local-resequencing revealed a novel 24bp indel polymorphism (henceforth referred to as Indel24) just under 2kb from the start of *SHROOM2*. This was subsequently genotyped in all 115 subjects in the eQTL analysis, with a minor allele frequency of 0.24. Due to its location within an ERV1 multiple repeat region, the exact origin of the indel polymorphism is ambiguous and may be arising at either chrX:9752561 or chrX:9752545 in hg19 (Figure 7.5).

Examination of publicly available sequencing data shows evidence for indel polymorphisms close to this site in two independent datasets – Phase 1 data from the 1000 genomes project and Complete Genomics (Drmanac *et al*, 2010).[†††] In the low coverage data (around 3x) from 1000 genomes (URL7.2), a 1bp indel was apparently detected at chrX:9752558. Complete Genomics provides higher coverage (around 80x) sequence for 69 individuals from a variety of human populations (URL7.3), and reports detection of a 24bp indel polymorphism at chrX:9752559. No other indels were detected in the Complete Genomics data within the ERV1 element or indeed anywhere in the extended *SHROOM2* promoter region.

Alignment of the alleles of the three indel polymorphisms is not straightforward as the structure of the repetitive sequence within and flanking the indels makes alignments in the region ambiguous. However it appears that all three indel events are consistent with a single site of origin (chrX:9752558-9752561) and that the two 24bp insertion alleles are almost identical except for a single nucleotide (Figure 7.6).

[†††] Examination of public datasets performed by Colin Semple, MRC Human Genetics Unit, IGMM.

```
          X:9752536                                                       X:9752584

RefSeq    CCACATCCCGCTGATTGGTCCATTT------------------------ACAGAGTGCTAATTGGTCCATTTT
Indel24   CCACATCCCGCTGATTGGTCCATTTTACAGAGTGCTGATTGGTCCATTTACAGAGTGCTAATTGGTCCATTTT


OR

RefSeq    CCACATCCC------------------------GCTGATTGGTCCATTTACAGAGTGCTAATTGGTCCATTTT
Indel24   CCACATCCCGCTGATTGGTCCATTTTACAGAGTGCTGATTGGTCCATTTACAGAGTGCTAATTGGTCCATTTT
```

Figure 7.5 Alignments demonstrating a novel polymorphic variant identified on targeted re-sequencing - a 24bp insertion at either X:9752561 or X:9752545, where the reference sequence lacks the insertion.

```
          X:9752536                                                         X:9752584

RefSeq    CCACATCCCGCTGATTGGTCCATTT------------------------ACAGAGTGCTAATTGGTCCATTTT
Indel24   CCACATCCCGCTGATTGGTCCATTTTACAGAGTGCTGATTGGTCCATTTACAGAGTGCTAATTGGTCCATTTT

RefSeq    CCACATCCCGCTGATTGGTCCA------------------------TTTACAGAGTGCTAATTGGTCCATTTT
CG_Ins    CCACATCCCGCTGATTGGTCCATTTTACAGAGTGCTAATTGGTCCATTTACAGAGTGCTAATTGGTCCATTTT

RefSeq    CCACATCCCGCTGATTGGTCCA-TTTACAGAGTGCTAATTGGTCCATTTT
1KG_Ins   CCACATCCCGCTGATTGGTCCATTTTACAGAGTGCTAATTGGTCCATTTT
```

Figure 7.6 Alignments of the three indel polymorphisms (Indel24: 24 bp insertion allele identified by our local-resequencing; CG_Ins: 24 bp insertion allele from Complete Genomics; 1KG_Ins: 1bp insertion allele from 1000 Genomes Phase 1 data). The sequence of Indel24 and CG_Ins is almost identical bar a single bp as highlighted in blue.

To investigate whether Indel24 is associated with *SHROOM2* expression, Indel24 genotypes were added to the genotypes for the 115 normal mucosa samples used for eQTL analysis and fine association mapping as described in Chapter 6. *SHROOM2* expression as quantified by qRT-PCR was used as the trait phenotype as it is thought to be a more accurate and sensitive measurement of transcript abundance than microarray signals.

A peak of association with *SHROOM2* expression was seen at X: 9,740,900 – 9,766,725 which encompass the tagging SNP rs5934683 (Figure 7.7). Distinctively, two variants were more significantly associated with *SHROOM2* expression than the tagging SNP, with p-values in the order of 1e-10. Closer examination reveals that the peak starts from the intergenic region between *GPR143*/*SHROOM2* and extends into the first intron of *SHROOM2*. The variants that were more significantly associated with *SHROOM2* than the tagging SNP were Indel24 and an intronic SNP rs5934685 (Figure 7.8) that had previously already been implicated in the fine-association mapping to *SHROOM2* as quantified on the microarrays (Table 6.5). In a similar fashion to the tagging SNP, the minor alleles for both these variants were associated with lower S*HROOM2* expression (Figure 7.9).

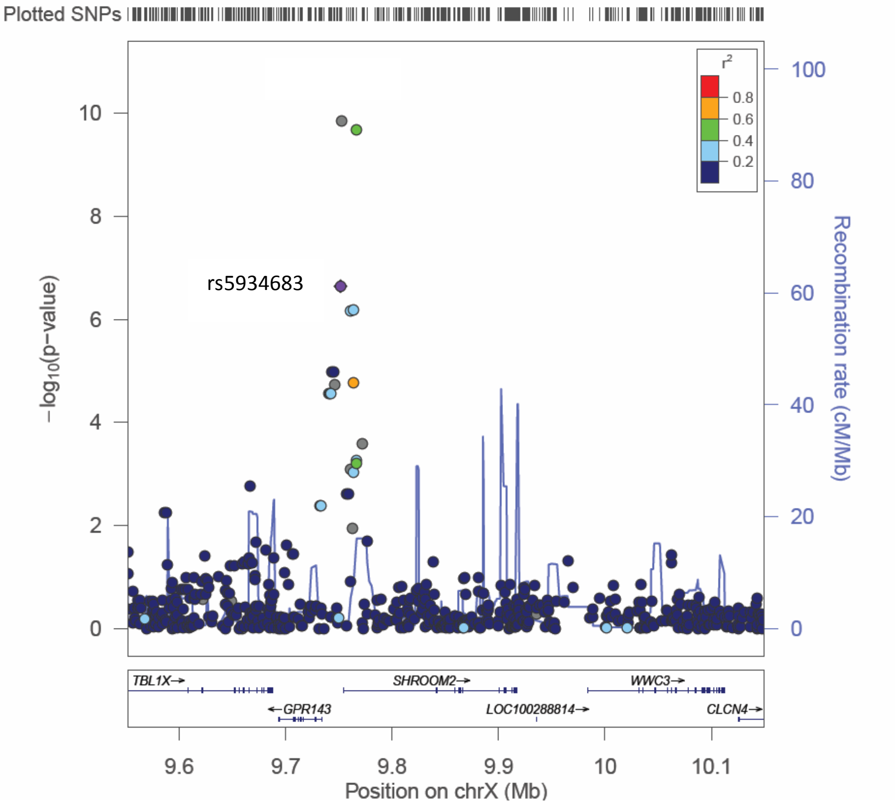


Figure 7.7 Manhattan plot displaying the strength of genetic association (-log$_{10}$ *p*-value) to *SHROOM2* expression versus chromosomal position, representing fine-mapping of a 600kb region surrounding the rs5934683 risk locus in 115 patients. The *p*-values were obtained by linear regression analysis with adjustment for age and gender. *SHROOM2* expression in the normal mucosa was measured by qRT-PCR, normalised to reference genes *EIF2B1*, *TBP* and *RPL30*. The peak of association maps to the tagging SNP and a 26kb surrounding region X: 9,740,900 – 9,766,725.

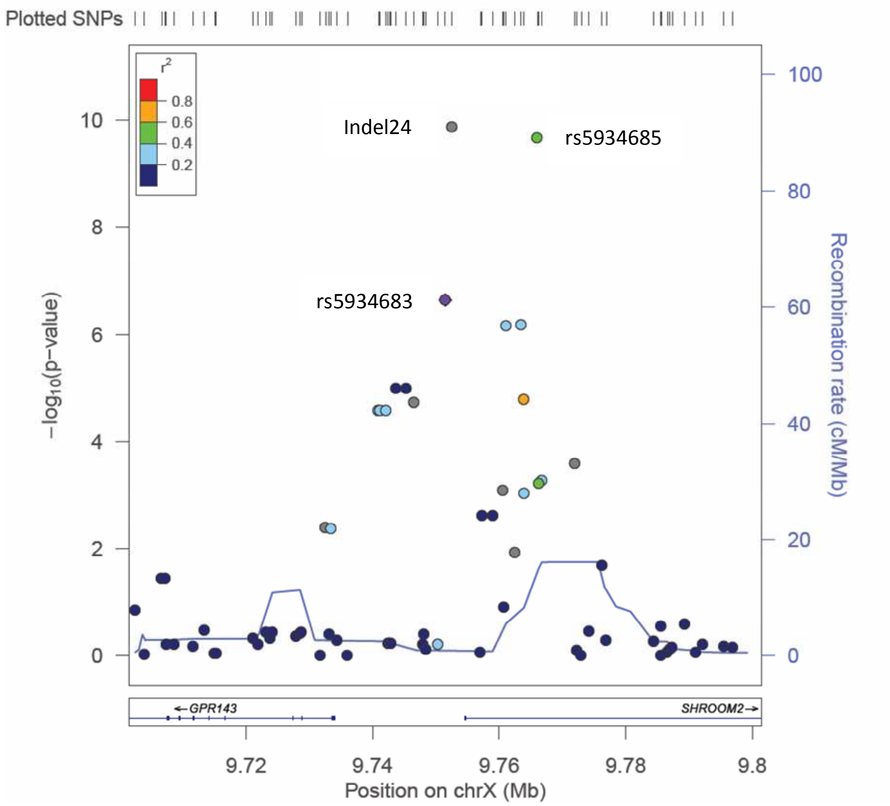
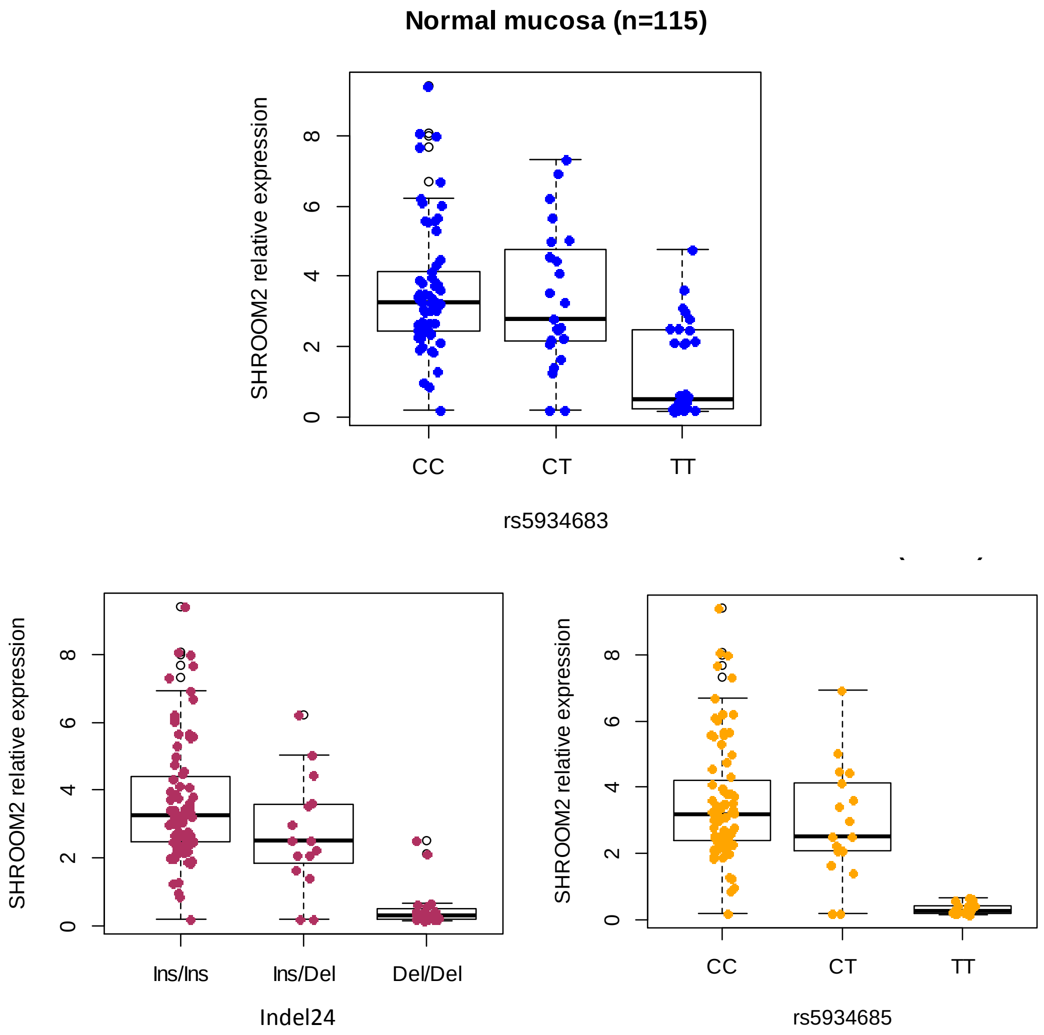


Figure 7.8 The peak of association with *SHROOM2* expression starts at the intergenic region between *GPR143/SHROOM2* and extends into the first intron of *SHROOM2*. There is no available linkage and recombination data in LocusZoom/HapMap CEU population (release 22) for Indel24.



Variant	Alleles	MAF	Estimate	*p*-value	Fold reduction
rs5934683	C/T	0.36 (T)	-1.056	2.59e-07	2.75 (95% CI, 1.96 – 4.36)
Indel24	Ins/Del	0.24 (Del)	-1.410	1.65e-10	6.72 (95% CI, 4.35 – 14.02)
rs5934685	C/T	0.23 (T)	-1.426	2.53e-10	10.5 (95% CI, 8.33 – 13.74)

Figure 7.9 Boxplots for the three variants most highly associated with *SHROOM2* expression. Estimated effect size and p-values are calculated from linear regression analysis with adjustment for age, gender and anatomical site. Fold reduction is the ratio of the expression means between the homozygotes of the major and the minor alleles.

The linkage disequilibrium structure of the peak region constructed from my sample set (n=115) shows that the two variants are in strong LD with each other and with the tagging SNP, suggesting that the top signals within the association peak are likely to be from a single association rather than two or three independent ones (Figure 7.10). To further understand which of these the functional variant is, linear regression modelling conditional on all three variants was performed on *SHROOM2* expression, as quantified by qRT-PCR and the Illumina HT12 microarray for comparison. The analysis performed on expression data derived from both methods indicates that the tagging SNP rs5934683 is not the causative variant, as the effect estimate and the test significance were markedly decreased when the two other variants were included in the model (Table 7.4). The interpretation of the test statistics for Indel24 and rs5934685 is not as straightforward; where expression was quantified by qRT-PCR, both Indel24 and rs5934685 bordered on significance, with Indel24 being the stronger signal both in terms of effect size and significance. Where expression was quantified by the HT12 microarray, rs5934685 appears to be the driver signal, attenuating the effect size and significance of Indel24. Although one may argue that the analysis based on qRT-PCR is more reliable as it has better detection sensitivity and larger dynamic range, it remains speculative at best to favour one variant over the other as the functional variant. The possibility of independent effects also cannot be excluded. Hence, follow-up with functional assays is critical to determine the functionality of these eQTL variants.

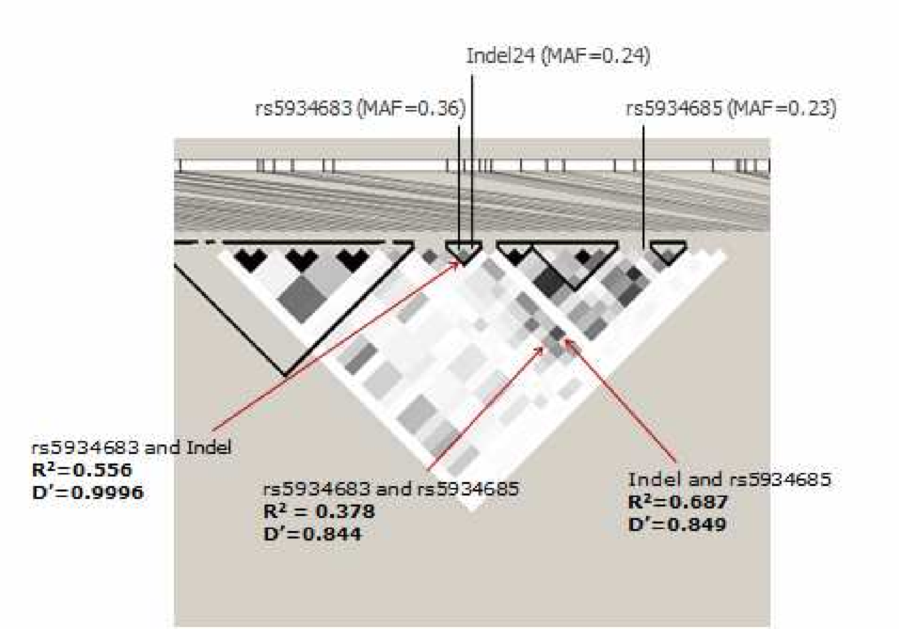


Figure 7.10 Linkage disequilibrium structure (r^2) surrounding the tagging SNP (rs5934683) and the 2 candidate functional variants (Indel24 and rs5934685) in my sample set (n=115). The D' reflects the frequency of co-inheritance of alleles, whereas the r^2 takes into further account the difference in the allele frequency.

	qRTPCR		HT12	
Variant	**Estimate**	***p*-value**	**Estimate**	***p*-value**
rs5934683	-0.0943	0.7454	-0.00297	0.908
Indel24	-0.8381	0.0592	-0.07217	0.066
rs5934685	-0.6831	0.0679	-0.14280	2.92e-05

Table 7.4 Linear regression for *SHROOM2* expression (as measured by qRTPCR or Illumina HT12 microarray), adjusted for age, gender, anatomical sampling site, the tagging SNP and the two putative causal variants. The estimate indicates the effect size.

7.3.3 Evidence of functionality for Indel24 and rs5934685 in ENCODE data

Indel24 is encompassed by an ERV repeat element (LTR12B; Family; ERV1; Class: LTR; Position: chrX:9752293-9752685) (Figure 7.11). Repeat elements are generally associated with increased indel rates (McDonald *et al*, 2011) and exapted ERV repeats have been reported to act as regulatory elements in human promoters (Cohen *et al*, 2009). More compellingly, Indel24 appears to lie within NF-YA and NF-YB transcription factor binding sites according to ENCODE ChIPseq data (Figure 7.11). As discussed previously (see 7.3.1), these two transcription factors bind cooperatively as two subunits of the trimeric NF-Y transcription factor complex, and often activate the transcription of cell cycle genes (Müller and Engeland, 2010). There is substantial published literature on NF-Y and it is known to have high affinity for the CCAAT box motif. Within known NF-Y binding sites, multiple CCAAT binding motifs are often found and the optimal spacing between them appears to be 24-53bp (Dolfini *et al*, 2009), which is similar to the spacing between the 3 CCAAT motifs found in this region (Figure 7.12). Remarkably, the 24bp insert contains a perfect match on the minus strand to the CCAAT box motif. This could, in theory, modulate NF-Y binding affinity either by creating/abolishing binding sites or by altering the spacing between them.

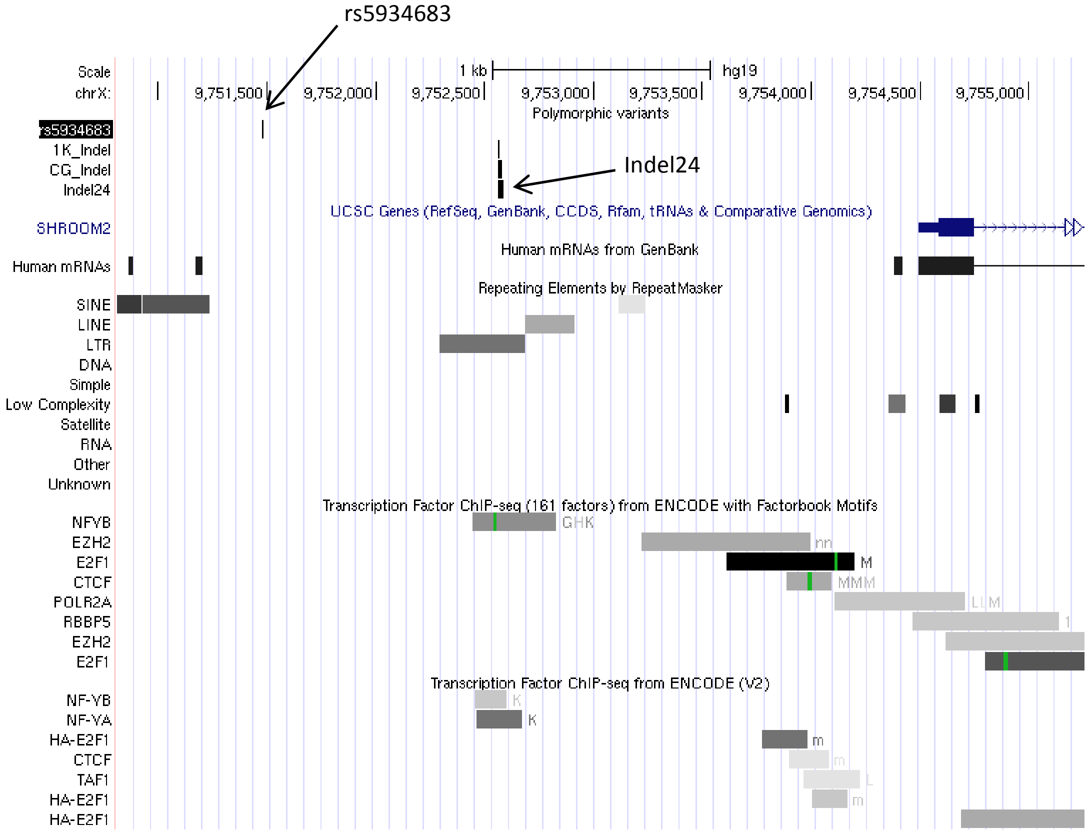


Figure 7.11 Genomic and regulatory landscape around Indel24, UCSC genome browser (1KG_Indel: 1bp indel from 1000 Genomes Phase 1 data; CG_Indel: 24 bp indel from Complete Genomics; Indel24: 24 bp indel as identified by our local-resequencing). The ERV1 repeat element is represented as LTR (long terminal repeat) in RepeatMasker track.

```
          X:9752536                                                     X:9752594

RefSeq    CCACATCCCGCTGATTGGTCCATTT------------------------ACAGAGTGCTAATTGGTCCATTTT
Indel24   CCACATCCCGCTGATTGGTCCATTTTACAGAGTGCTGATTGGTCCATTTACAGAGTGCTAATTGGTCCATTTT


          X:9752595                                                      X:9752657

RefSeq    ACAAACCTCTAGCTAGCCACAGAGCGCTGATTGGTGCATTTTACAATCCTCTTGTAAGACAGAAAAATTCTCG
Indel24   ACAAACCTCTAGCTAGCCACAGAGCGCTGATTGGTGCATTTTACAATCCTCTTGTAAGACAGAAAAATTCTCG
```

Figure 7.12 Three CCAAT box motifs on the minus strand (appearing as ATTGG on the plus strand; pink highlight) are found within the reference sequence of the NFY-B binding site at Xp22.2, with 23bp and 42bp spacing between them. The 24bp insert alters the spacing to 47bp and 42bp, or donates a fourth CCAAT box motif (green highlight), with spacing between the motifs of 24bp, 23bp and 42bp.

The other SNP that is a putative causal variant, rs5934685, is a *SHROOM2* intronic variant at chrX:9766019. There is no evidence in the literature to support eQTL activity for it, and according to ENCODE it is not associated with DNase hypersensitivity, transcription factor binding, histone modifications or methylation. The closest transcription factor binding site is situated 324bp upstream from the SNP, which binds to YY1(Ying Yang 1) with a cluster score of 180/1000 in H1-hESC and NT2-D1 cell lines (Figure 7.13). YY1 a multifunctional zinc-finger transcription factor that has been associated with cellular proliferation and resistance to apoptotic stimuli, and is known to be over-expressed in colorectal cancer (Chinnapan *et al*, 2009). There is also a conserved transcription factor binding site 196bp upstream of the SNP, which is predicted to bind to VSX2 (visual system homeobox 2). This is of possible relevance as VSX2 was originally described as a retina-specific transcription factor, with mutations associated with microphthalmia, cataracts and iris abnormalities (NCBI gene; URL7.4). It was mutated to create the first mouse model of retinoblastoma (Zhang *et al*, 2004) and has been reported as a novel biomarker for CRC (Mori *et al*, 2011).

In summary, local targeted resequencing and fine-mapping strategies have identified 2 putative causative variants for the eQTL activity observed at the rs5934683 locus. One of these variants, Indel24, is a novel indel polymorphism with in-silico evidence of NF-Y transcription factor binding properties. Statistical modelling of SHROOM2 expression accounting for rs5934683 and the 2 candidate variants indicates that rs5934683 is a tagging proxy, but is inconclusive in determining which of the 2 candidates is the functional variant driving the eQTL association. Further evaluation with in-vitro regulatory assays will be necessary to validate these observations and provide insight into the mechanisms underlying causation.

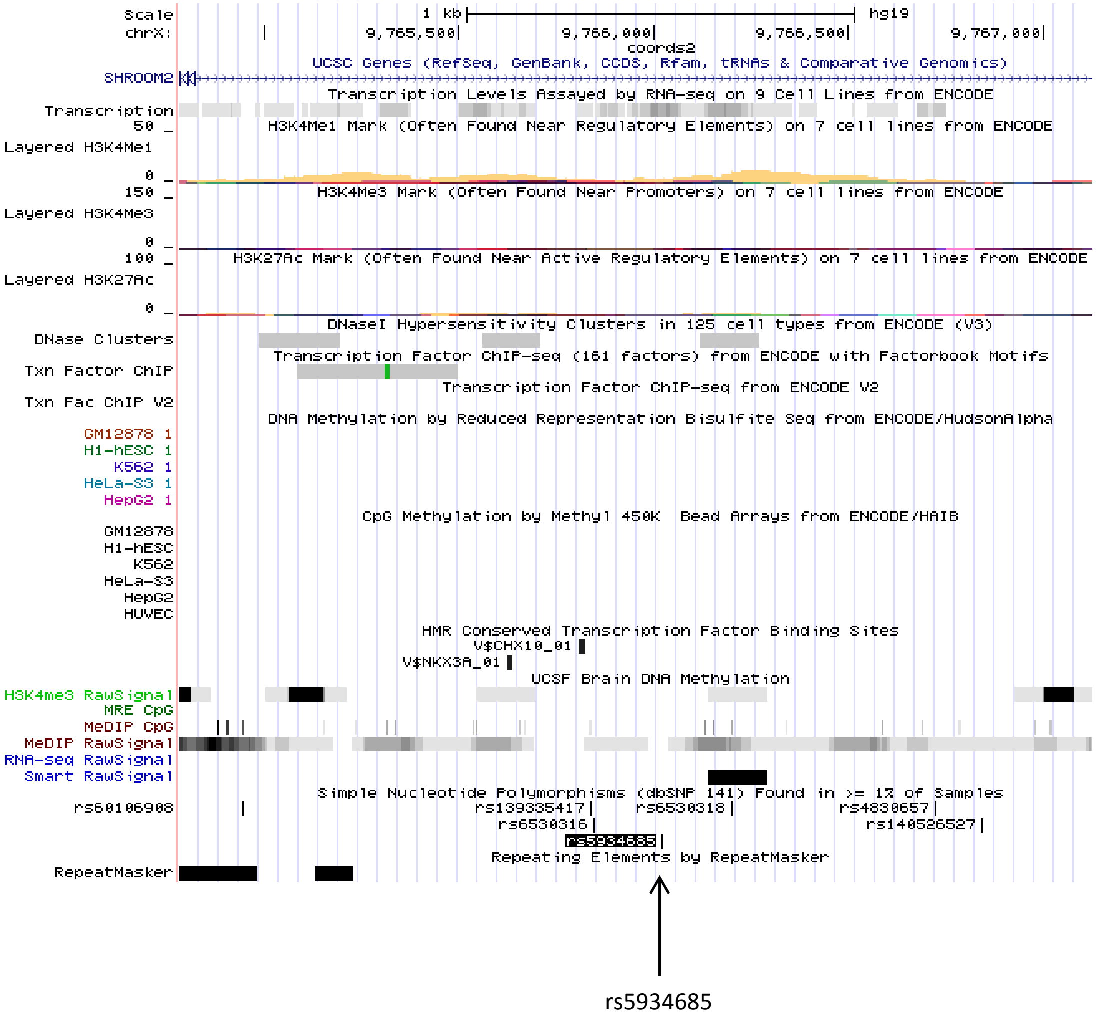

Scale
1 kb
hg19
chrX:
9,765,500|
9,766,000|
9,766,500|
9,767,000|
coords2
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
SHROOM2
Transcription Levels Assayed by RNA-seq on 9 Cell Lines from ENCODE
Transcription
50 _
H3K4Me1 Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K4Me1
0 _
150 _
H3K4Me3 Mark (Often Found Near Promoters) on 7 cell lines from ENCODE
Layered H3K4Me3
0 _
100 _
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
0 _
DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V3)
DNase Clusters
Transcription Factor ChIP-seq (161 factors) from ENCODE with Factorbook Motifs
Txn Factor ChIP
Transcription Factor ChIP-seq from ENCODE V2
Txn Fac ChIP V2
DNA Methylation by Reduced Representation Bisulfite Seq from ENCODE/HudsonAlpha
GM12878 1
H1-hESC 1
K562 1
HeLa-S3 1
HepG2 1
CpG Methylation by Methyl 450K Bead Arrays from ENCODE/HAIB
GM12878
H1-hESC
K562
HeLa-S3
HepG2
HUVEC
HMR Conserved Transcription Factor Binding Sites
V$CHX10_01
V$NKX3A_01
UCSF Brain DNA Methylation
H3K4me3 RawSignal
MRE CpG
MeDIP CpG
MeDIP RawSignal
RNA-seq RawSignal
Smart RawSignal
Simple Nucleotide Polymorphisms (dbSNP 141) Found in >= 1% of Samples
rs60106908
rs139335417
rs6530318
rs4830657
rs6530316
rs140526527
rs5934685
Repeating Elements by RepeatMasker
RepeatMasker
rs5934685

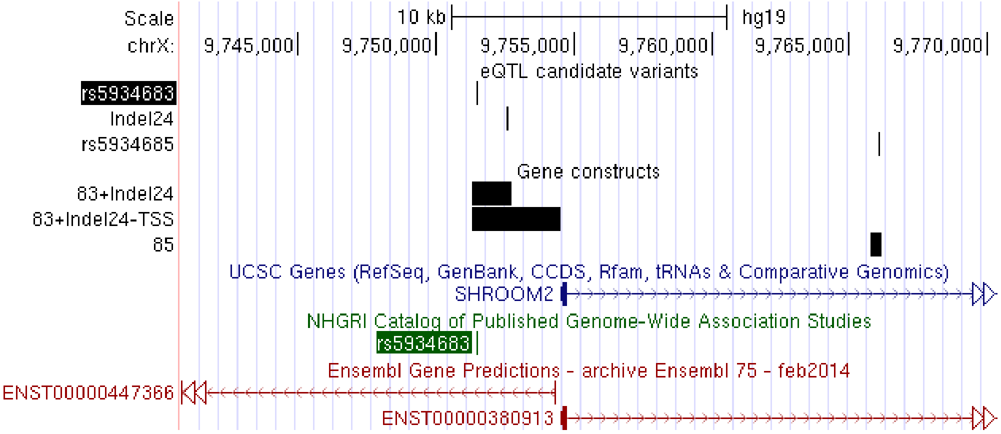
Figure 7.13 The genomic and regulatory landscape around r5934685 (UCSC browser).

7.3.4 Transcriptional activity assays on putative causal variants implicates Indel24 as the functional variant

To investigate whether the candidate variants possess allele-specific regulatory effects, gene elements containing the different alleles of the 3 candidate variants were cloned from human genomic DNA into luciferase reporter vectors (Figure 7.14).

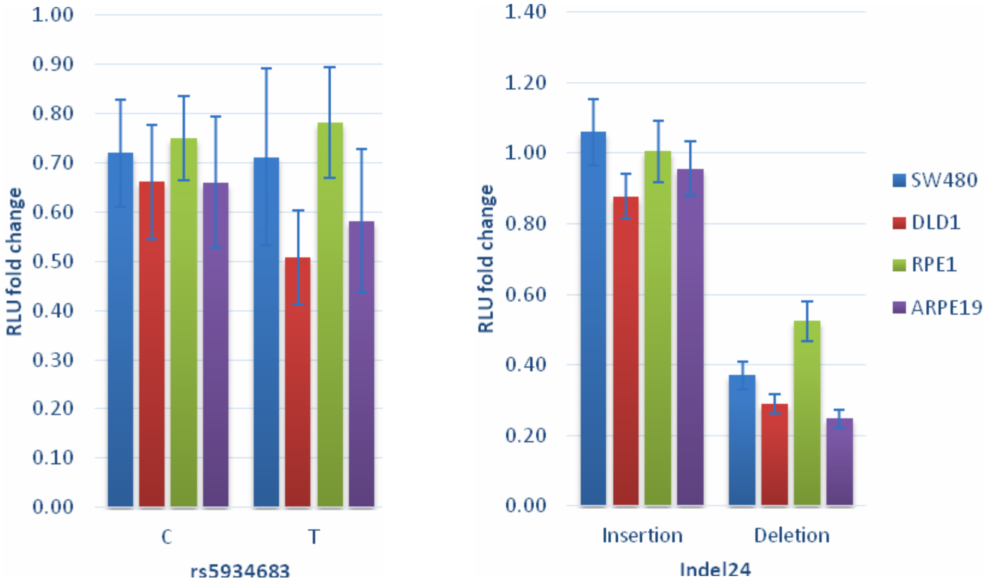
Given the proximity of Indel24 to rs5934683, a 1449bp gene construct (referred to as 83+Indel24) containing both variants was tested, with each of the four possible haplotypes cloned into a basic transcriptional reporter vector (pGL2). All four test constructs were transfected into two CRC cell lines (SW480 and DLD1) and two retinal pigment epithelial cell lines (RPE1 and ARPE1).[‡‡‡] Indel24 shows a highly significant allele-specific differential effect on luciferase activity in all four cell lines, with the deletion allele showing a stark reduction in transcriptional activity (Figure 7.15). This is in contrast to the lack of effect between the different alleles of rs5934683. There was also no statistical interaction between rs5934683 and Indel24 in all cell lines, which indicates that the regulatory differences seen with the Indel24 alleles are independent of the rs5934683 variant.

[‡‡‡] The extra-colonic function of *SHROOM2* was investigated by PhD student Andrew McBride and the reporter assays in non-CRC cell lines were performed by him. This data has been included here for completeness as it formed part of a figure that has been submitted for publication.



Gene constructs	Location	Size (bp)	Features	Alleles/ Haplotypes	Vector
83+Indel24	chrX: 9751297 - 9752746	1449	Encompasses rs5934683 and Indel24	C; Insertion C; Deletion T; Insertion T; Deletion	pGL2
83+Indel24-TSS	chrX: 9751297 - 9754496	3199	Encompasses rs5934683 and Indel24, extends to the start of *SHROOM2*	C; Insertion C; Deletion T; Insertion T; Deletion	pGL2
85	chrX: 9765761 - 9766181	421	Encompasses rs5934685	C T	pGL2 pGL4

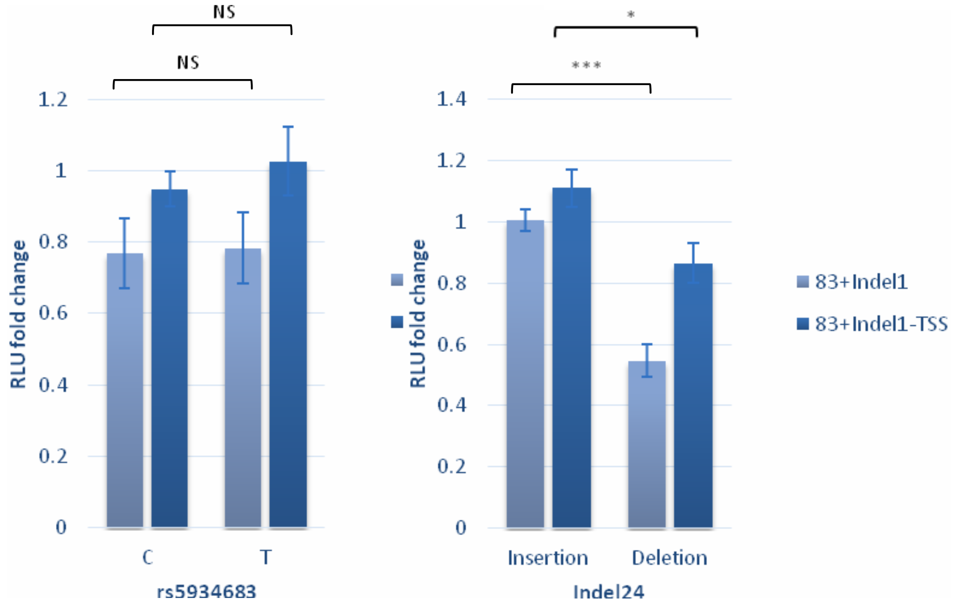
Figure 7.14 3 gene constructs containing the 3 eQTL candidate variants (in combination or individually, see table) were generated from human genomic DNA. The 83+Indel24 and 83+Indel24-TSS elements were cloned into pGL2 basic transcriptional vectors, whereas the 85 gene element was cloned into pGL2 and pGL4 as well to maximise detection of enhancer activity.



Variant	Cell line	*p*-value	Fold change
rs5934683	SW480	0.97	-
rs5934683	DLD1	0.31	-
rs5934683	RPE1	0.81	-
rs5934683	ARPE19	0.67	-
Indel24	SW480	2.14e-05***	0.35 (95% CI, 0.26 - 0.47)
Indel24	DLD1	1.46e-08***	0.33 (95% CI, 0.26 - 0.42)
Indel24	RPE1	1.65e-04***	0.51 (95% CI, 0.38 - 0.68)
Indel24	ARPE19	6.42e-07***	0.26 (95% CI, 0.21 - 0.32)
rs5934683*Indel24	SW480	0.24	-
rs5934683*Indel24	DLD1	0.12	-
rs5934683*Indel24	RPE1	0.86	-
rs5934683*Indel24	ARPE19	0.91	-

Figure 7.15 Luciferase reporter assays indicating transcriptional activity of the 83+Indel24 (pGL2) gene construct. The constructs for each allele were transfected into 4 cell lines (SW480 and DLD1 are CRC cell lines, RPE1 and ARPE19 are retinal pigment epithelial cell lines), and the experiment was replicated at least 4 times in each cell line. Error bars=SEM. The allele-specific reporter activity of the rs5934683 SNP and the Indel24 variant was analysed separately and together to assess possible interactions. Table shows ANOVA *p*-values; where significant, effect sizes were calculated.

To increase the evidence that the transcriptional activity seen with 83+Indel24 is reflective of the in-vivo regulation of *SHROOM2*, the gene element was extended to include the core promoter region/transcription start site (TSS) of *SHROOM2*. The experiment was repeated in SW480 with the larger 83+Indel24-TSS gene construct, with the shorter 83+Indel24 as a positive control. Although overall activity was attenuated and the differential effect was reduced in the larger construct (effect size of 0.78 (95% CI, 0.63 - 0.95)), there was still a significant allele-specific effect seen with Indel24 that mirrors that seen in the shorter construct (Figure 7.16). Again, this effect was not observed between the different alleles of rs5934683.

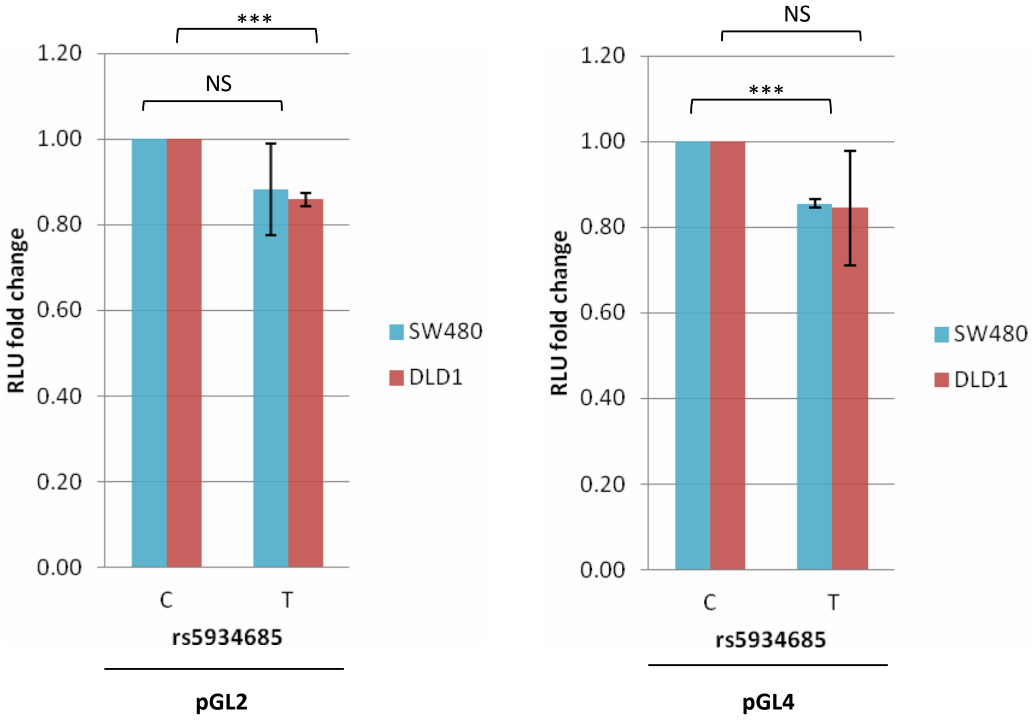


Variant	Gene element	*p*-value	Effect size (ratio)
rs5934683	**83+Indel24**	0.92	-
rs5934683	**83+Indel24-TSS**	0.49	-
Indel24	**83+Indel24**	4.72e-06***	0.54 (95% CI, 0.43 - 0.67)
Indel24	**83+Indel24-TSS**	0.016*	0.78 (95% CI, 0.63 - 0.95)

Figure 7.16 Luciferase reporter activity for the 83+Indel24 (pGL2) and the longer 83+Indel24-TSS (pGL2) gene construct. The constructs for each allele was transfected into CRC cell line SW480 and the experiment replicated 4 times. Error bars=SEM. Table presents unpaired Student t-test *p*-values; where significant, effect sizes were calculated.

To test if the intronic variant rs5934685 has a regulatory effect, a 421bp gene construct (referred to as 85; Figure 7.14) encompassing the rs5934685 SNP was cloned into pGL2 as well as pGL4 to maximise detection of enhancer activity. The test constructs were transfected into SW480 and DLD1, and appeared to exhibit a small degree of allele-specific differential effects on reporter activity (Figure 7.17). When the 85 gene element was cloned into pGL2, there appeared to be a reduction in reporter activity with the T allele (effect size of 0.86 (CI 95%, 0.82 – 0.90)), but this effect was only observed in one of the two cell lines, DLD1. The same gene element in the enhancer reporter pGL4 also showed a reduction in reporter activity with the T allele (effect size of 0.86 (CI 95%, 0.84 – 0.88)). Again this was only seen in one of the two tested cell lines, SW480. Although these effects were significant, they were relatively small, with a 14% average reduction in transcriptional activity from the C allele.

In conclusion, this demonstrates that Indel24 has allele-specific regulatory function, whereas this is not readily apparent of rs5934683. These results concur with ENCODE data where Indel24 is located in known regulatory elements. On the other hand, there is some evidence that rs5934685 may have regulatory properties. Although the allele-specific transcriptional effect of the 85 gene construct is modest in comparison to that seen with the 83+Indel24 construct, the effect sizes and confidence intervals of the 85 construct are comparable to that of the longer 83+Indel24-TSS construct. Overall, these results suggest that Indel24 is most likely to be the causative variant driving the *SHROOM2* eQTL association, with rs5934683 tagging the locus signal. However, rs5934685 cannot be ruled out as an independent causative variant contributing an independent regulatory effect.



Vector	Cell line	*p*-value	Effect size (ratio)
pGL2	**SW480**	0.31	-
pGL2	**DLD1**	9.18e-05	0.86 (95% CI, 0.82 – 0.90)
pGL4	**SW480**	7.48e-06	0.86 (95% CI, 0.84 – 0.88)
pGL4	**DLD1**	0.29	-

Figure 7.17 Luciferase reporter activity for the 85(pGL2) construct and 85(pGL4) construct. The constructs for each allele was transfected into two CRC cell lines (SW480 and DLD1) and each experiment replicated 4 times. Error bars=SEM. Table presents unpaired Student t-test *p*-values; where significant, effect sizes were calculated.

7.3.5 Indel24 polymorphism alters transcriptional activity of *SHROOM2* by influencing NF-Y binding affinities

The Indel24 polymorphism, defined by the presence or absence of a 24bp gene element at chrX: 9752561, appears to have allele-specific transcriptional properties that may explain the *SHROOM2* eQTL association in colonic normal mucosa. As discussed above, this region has been found to bind NFY transcription factor subunits A and B in the ENCODE project ChIP-seq data. This suggests that Indel24 may be modulating transcription of *SHROOM2* by altering the DNA-binding affinity of NF-Y.

To confirm the role of NF-Y in Indel24-mediated transcriptional activity, siRNA knockdown of NFY-A and NFY-B was carried out in MCF7 cell line, which was co-transfected with the 83-Indel24 (pGL2) construct containing the insertion allele. Two siRNAs were always used for each gene to enable detection of non-specific or off-target effects. *NF-YA* and *NF-YB* mRNA and protein levels were assessed to ensure effective knockdown (Figure 7.18). There was a significant decrease of 30-40% in the associated reporter activity upon *NF-YA* or *NF-YB* depletion (Figure 7.19). This finding is similar between the two NF-Y subunits, which fits in with the knowledge that all three subunits of the heterotrimeric complex are required for DNA binding. This effect is recapitulated with endogenous *SHROOM2*, whereby siRNA knockdown of *NF-YA* or *NF-YB* in DLD1, SW480 and RPE1 cell lines are associated with a significant decrease in *SHROOM2* mRNA levels (Figure 7.20 and 7.21). *CCNB1* (Cyclin B1) is measured as a positive control as it has a well-characterised *NF-Y* promoter (Mani *et al*, 2001) in various cell types including colorectal cancer cell lines (Jürchott *et al*, 2010).

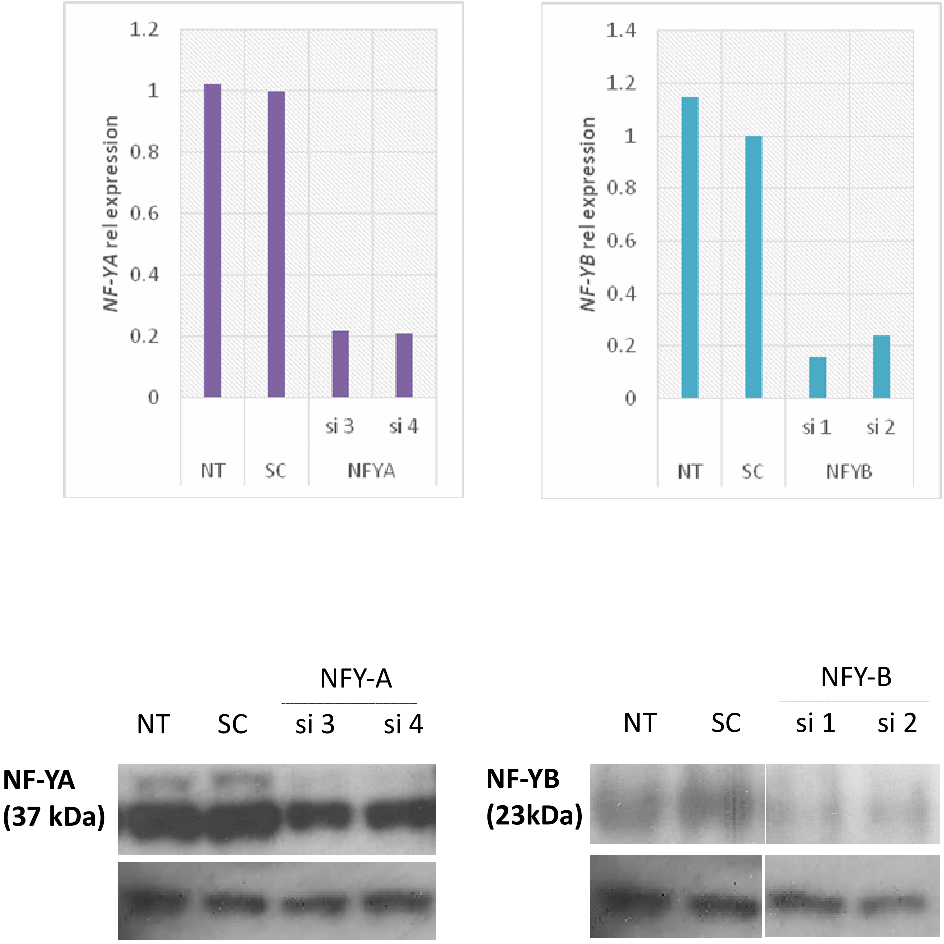
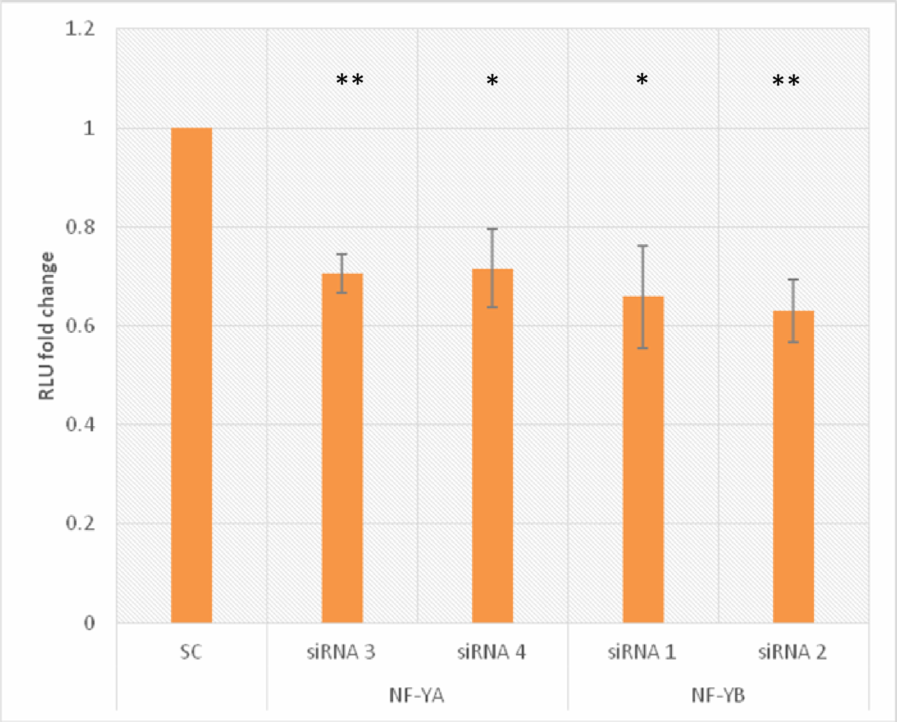
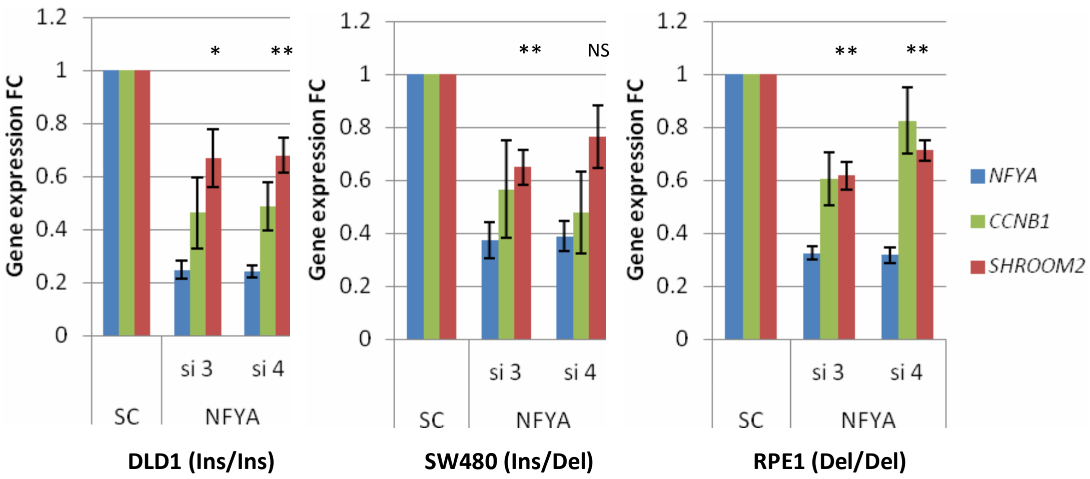


Figure 7.18. siRNA knock-down of *NF-YA* and *NF-YB* in MCF7 as assessed by qRT-PCR (top panel) and Western Blotting (bottom panel). Graphs and blots shown are representative of 3 replicates. mRNA expression levels of *NF-YA* and *NF-YB* were normalised to reference gene *TBP*, whereas β-actin was used as the loading control for Western Blots. (NT=non-treated, SC=scrambled control, si=siRNA)



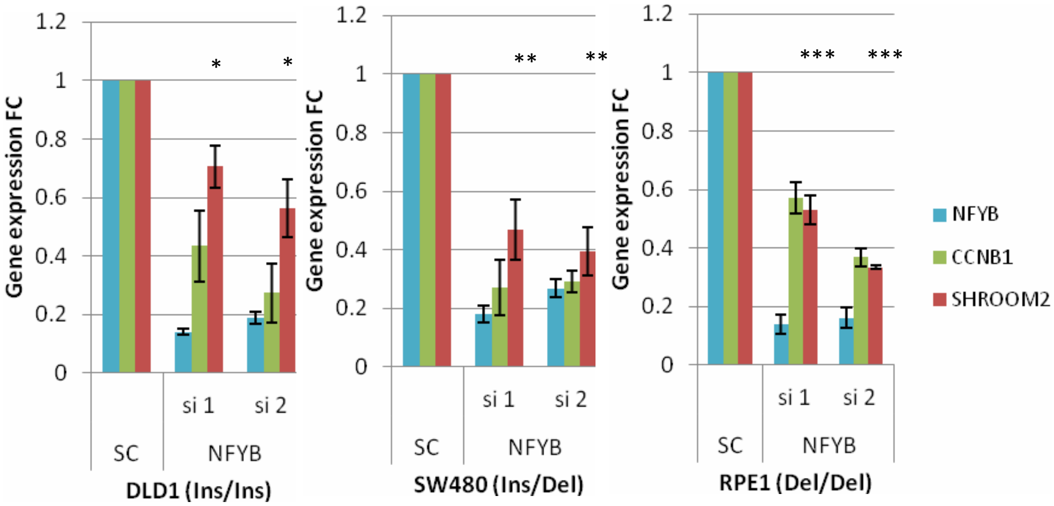
Gene	siRNA	*p*-val	Fold change
NF-YA	**3**	0.0017**	0.71 (95% CI, 0.60 - 0.81)
NF-YA	**4**	0.0231*	0.72 (95% CI, 0.50 - 0.94)
NF-YB	**1**	0.0293*	0.66 (95% CI, 0.37 - 0.94)
NF-YB	**2**	0.0043**	0.63 (95% CI, 0.46 - 0.81)

Figure 7.19 Luciferase reporter activity of the 83-Indel24 (pGL2) insertion allele, when *NF-YA* or *NF-YB* expression was knocked down with siRNA. MCF-7 cell line was used for the co-transfection. *p*-values reported are of unpaired Student t-tests comparing each siRNA was to the scrambled control (SC). Error bars=SEM. Experiments replicated 3 times.



Cell line	siRNA	*p*-val	Fold change
DLD1	**3**	0.0389*	0.67 (95% CI, 0.36 - 0.97)
DLD1	**4**	0.0079**	0.68 (95% CI, 0.50 - 0.86)
SW480	**3**	0.0059**	0.65 (95% CI, 0.47 - 0.83)
SW480	**4**	0.1170	-
RPE1	**3**	0.0018**	0.62 (95% CI, 0.48 - 0.76)
RPE1	**4**	0.0016**	0.71 (95% CI, 0.61 - 0.82)

Figure 7.20 Reduction in *SHROOM2* expression observed when *NF-YA* expression was knocked down with two different targeting siRNAs (si3 and si4) in three different cell lines (DLD1, SW480 and RPE1). The Indel24 genotypes of the cell lines are represented as Ins/Ins (homozygote for insertion), Ins/Del (heterozygote) and Del/Del (homozygote for deletion). Expression quantified by qRT-PCR, normalised to reference genes *TBP* or *RPL30.* Error bars=SEM. Experiment replicated 3 times. *p*-values reported are of unpaired Student t-tests comparing *SHROOM2* expression between each of the siRNAs and the scrambled control (SC).

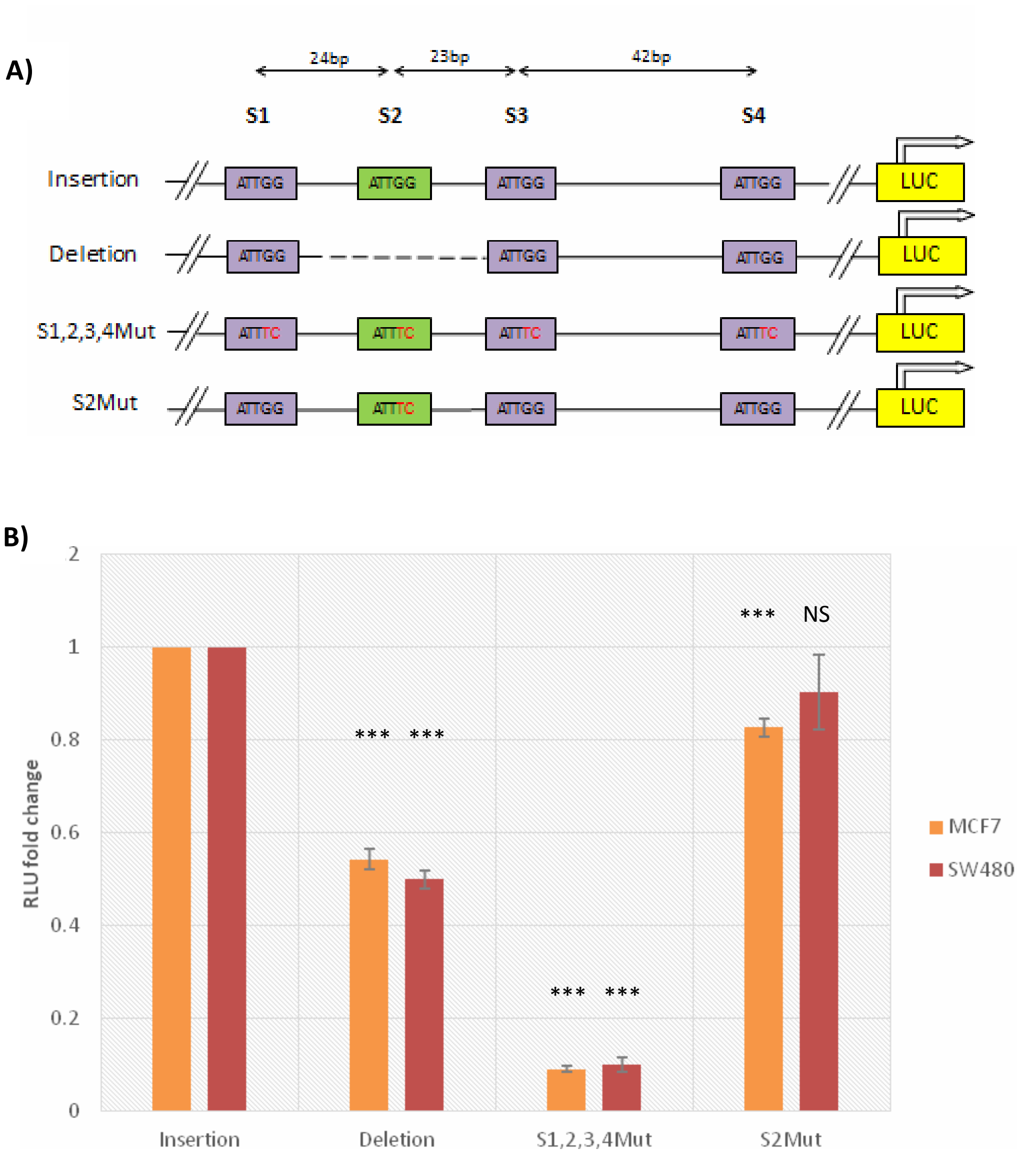


Cell line	siRNA	*p*-val	Fold change
DLD1	**1**	0.0152*	0.71 (95% CI, 0.51 - 0.91)
DLD1	**2**	0.0115*	0.56 (95% CI, 0.29 - 0.84)
SW480	**1**	0.0066**	0.47 (95% CI, 0.18 - 0.75)
SW480	**2**	0.0018**	0.39 (95% CI, 0.16 - 0.62)
RPE1	**1**	0.0007***	0.53 (95% CI, 0.39 - 0.67)
RPE1	**2**	<0.0001***	0.33 (95% CI, 0.31 - 0.35)

Figure 7.21 Reduction in *SHROOM2* expression observed when *NF-YB* expression was knocked down with two different targeting siRNAs (si1 and si2) in three different cell lines (DLD1, SW480 and RPE1). The Indel24 genotypes of the cell lines are represented as Ins/Ins (homozygote for insertion), Ins/Del (heterozygote) and Del/Del (homozygote for deletion). Expression quantified by qRT-PCR, normalised to reference genes *TBP* or *RPL30*. Error bars=SEM. Experiment replicated 3 times. *p*-values reported are of unpaired Student t-tests comparing *SHROOM2* expression between each of the siRNAs and the scrambled control (SC).

The reduction in 83-Indel24 luciferase reporter activity and endogenous *SHROOM2* with *NF-YA* or *NF-YB* knockdown strongly suggests that NF-Y plays a regulatory role in *SHROOM2* transcription by binding to the DNA region encompassing the Indel24 eQTL. This is entirely plausible, as NF-Y is the major CCAAT-binding factor (Testa *et al*, 2005; Ceribelli *et al*, 2008), and there is an ATTGG motif (CCAAT on the minus strand) within the 24bp insertion element with three other ATTGG motifs in very close proximity (Figure 7.16). Indeed, CCAAT box motifs can be found in promoter regions in either the CCAAT or ATTGG orientation, and multiple CCAAT box motifs have been observed for NF-Y promoters (Dolfini *et al*, 2009). However, one potentially confounding caveat to this is that there are two other CCAAT box motifs within the reporter gene construct 83-Indel24 at chrX:9751516 (1kb upstream of Indel24) and chrX:9752716 (150kb downstream of Indel24) that could be contributing to NF-Y driven reporter activity. To clarify whether the ATTGG motifs at the Indel24 site are the functional motifs, these motifs within the 83-Indel24 gene construct were mutated to ATTTC (Figure 7.22A), which is predicted by JASPAR to have very low NF-Y binding properties. The reporter assays performed on the S1,2,3,4 mutant construct demonstrates that transcriptional activity was dramatically reduced by ~90% (Figure 7.22B), strongly suggesting that the ATTGG motifs at the Indel24 locus are the functional motifs, with NF-Y binding at the other two farther sites much less likely.

As alluded to previously, it is known that the spacing between motifs in multiple CCAAT binding sites is important. To distinguish whether the insertion element of Indel24 improves transcriptional activity by increasing the spacing between S1-S3, or by donating an extra binding site in the form of S2, S2 was mutated in the reporter construct 83-Indel24 (Figure 7.22A). S2Mut did not reproduce the impact of the Deletion allele; its reporter activity was only minimally reduced from the Insertion allele (Figure 7.22B). This strongly indicates that Indel24 modifies NF-Y binding by altering the spacing between S1-S3 and not by increasing the number of binding sites.



Gene construct	Cell line	*p*-value	Fold change
Deletion	MCF7	3.20e-05***	0.54 (95% CI, 0.48 - 0.60)
Deletion	SW480	1.72e-05***	0.50 (95% CI, 0.44 - 0.56)
S1,2,3,4Mut	MCF7	1.90e-08***	0.09 (95% CI, 0.07 - 0.11)
S1,2,3,4Mut	SW480	6.30e-07***	0.10 (95% CI, 0.06 - 0.15)
S2Mut	MCF7	8.92e-04***	0.83 (95% CI, 0.77 - 0.88)
S2Mut	SW480	0.30	-

Figure 7.22 A) Mutations to the CCAAT box motif (ATTGG on the plus strand) were introduced to the insertion allele of the 83-Indel24 gene construct. B) Luciferase reporter activity for both alleles of the 83-Indel24 gene construct and the mutant constructs. Error bars=SEM, experiment was replicated three times. Table shows *p*-values for Student unpaired t-tests comparing the means of each construct to the Insertion allele.

7.3.6 Case-control studies demonstrates Indel24 as the functional variant for CRC risk

A CRC case-control logistic regression analysis was performed for all variants across the rs5934683 risk locus in 687 cases and 873 controls from the SOCCS (Scottish Colorectal Cancer Susceptibility) study. Of the three candidate variants, Indel24 appears to be most significantly associated with risk (*p* =0.05), with the Insertion allele conferring an OR of 0.86 (95% CI, 0.75 - 1.00). The tagging SNP rs5934683 and the other candidate variant rs5934685 did not reach significance (Table 7.5). When the model was conditioned on all three variants, Indel24 again stood out as the only significant variant (*p*=0.02) with an OR of 0.66 (95% CI, 0.46 - 0.94).

Subsequently, the effect of Indel24 on risk was validated in the larger dataset derived from samples from across Scotland, England and Croatia (8368 cases and 6327 controls). Indel24 is more significantly associated with the disease phenotype (*p*=0.03, OR=0.92) compared to rs5934683 (*p* =0.47, OR=0.98) (Table 7.6).

Overall, these results strongly suggest that Indel24 is indeed the causative variant for CRC risk within this eQTL locus.

	Variant	*p*-value	Odds Ratio
Individual analysis	rs5934683	0.66	0.97 (95% CI, 0.86 - 1.10)
	Indel24	0.05	0.86 (95% CI, 0.75 - 1.00)
	rs5934685	0.11	0.88 (95% CI, 0.76 - 1.03)
Conditional on all three putative variants	rs5934683	0.16	1.14 (95% CI, 0.95 - 1.37)
	Indel24	0.02	0.66 (95% CI, 0.46 - 0.94)
	rs5934685	0.27	1.22 (95% CI, 0.86 - 1.73)

Table 7.5 Association between putative causal variants and the risk of CRC in SOCCS study (687 cases and 873 controls from the Scottish population). *p*-values and odds ratios were derived from the logistic regression model adjusted for age and gender. The top panel shows the results for the individual analysis of each of the variants, whereas the bottom panel shows the results for the conditional modelling when all three variants were included as co-variates.

	Variant	*p*-value	Odds Ratio
Individual analysis	rs5934683	0.47	0.98 (95% CI, 0.93 - 1.04)
	Indel24	0.03	0.92 (95% CI, 0.86 – 0.99)

Table 7.6 The CRC association of Indel24 compared to the tagging SNP rs5934683 in 8368 cases and 6327 controls from Scottish, English and Croatian populations. Each variant was analysed separately, with *p*-values and odds ratios derived from conditional logistic regression adjusted for gender, age and country.

7.4 Discussion

Whole-genome gene expression profiling has identified the colorectal cancer (CRC) risk locus at Xp22.2 to be associated with the colonic mucosa expression levels of a neighbouring gene *SHROOM2*, providing a functional mechanism for this low-risk genetic locus. Interrogation of the locus with targeted re-sequencing and fine-mapping has identified two putative causal variants that appear to drive the association with *SHROOM2* expression - a novel genetic control element (Indel24) at -2203 and rs5934685 within intron 1. Both are significantly more associated with expression and colorectal cancer risk than the tagging SNP rs5934683. Conditional analysis is suggestive that Indel24 is the driver signal in a case-control study, but this was not conclusive in the eQTL analysis. Hence, it is crucial that these empirical observations are analysed in context with functional studies, as this will help to demonstrate the mechanism underlying the eQTL association, and offer insight into the aetiology of inherited colorectal susceptibility.

The expression of a gene can be influenced in several ways by a genetic variant, be it by influencing epigenetic mechanisms such as methylation, altering transcriptional activity, or modifying the stability of transcripts to degradation. The location of the candidate causal variants at the 5' end of *SHROOM2* is suggestive of an influence on transcriptional activity or possibly promoter methylation. Indeed, luciferase reporter assays provided evidence that Indel24 exhibits strong allele-specific differences in transcriptional activity, whereas the other putative causal candidate rs5934685 had only a weak effect, if any. These assays also confirm the lack of effect of rs5934683, confirming its role as a tagging SNP. Conversely, whole genome and localised methylation analysis performed collaboratively with other members of the group did not reveal any evidence of differential methylation in region of the tagging SNP and Indel24.

The finding that Indel24 has allele-specific regulatory control of transcription is consistent with cell line ChIP-seq data from ENCODE which indicates that Indel24 resides in an NF-Y transcription factor binding site within an ERV1 repeat element. Repeat elements are generally associated with increased indel rates (McDonald *et al*,

2011) and exapted ERV repeats have been reported to act as regulatory elements in human promoters (Cohen *et al*, 2009). This repeat may therefore be the source of indel polymorphisms in this region, as well as provide a mechanism for the observed eQTL effects. Examination of FANTOM5 data, which provides deeply sequenced CAGE data over ~2000 human cell types suggests that the ERV1 element is not transcribed and is not an alternative *SHROOM2* promoter (Semple C, pers. comm.). Given the distance of this ERV1 repeat from the *SHROOM2* TSS (~1.5 Kb) it would seem more reasonable to think of this region as a distal promoter element rather than an enhancer. This is consistent with ENCODE data which does not show the characteristic chromatin signature of an enhancer.

Further support for Indel24 as the functional variant was provided by in-silico analysis which shows that Indel24 harbours an NF-Y binding motif (CCAAT box), with multiple other CCAAT motifs flanking the Indel24 sequence. Depletion of the NF-Y subunits as well as mutation of the CCAAT binding motifs was associated with a reduction in Indel24 reporter activity, implicating NF-Y as the transcription factor that is interacting with Indel24 within the ERV1 repeat element to modify *SHROOM2* levels. This makes biological sense, as there have been reports of intergenic ERV repeats recruiting NF-Y in adult erythroid cells to assemble a complex including RNA polymerase II and thereby affect downstream transcription of genes (Pi *et al*, 2010). Furthermore, the *SHROOM2* promoter has two binding sites for the E2F1 transcription factor near the *SHROOM2* TSS and studies have indicated that E2F1 and NF-YA can bind to promoters cooperatively to activate transcription (Ru *et al*, 2006). Biochemical approaches can be adopted to consolidate these findings, for example ChIP (chromatin immunoprecipitation) and EMSAs (Electrophoretic mobility shift assays) with NF-Y and E2F1 antibodies would be useful follow-up studies that can demonstrate endogenous and in-vitro DNA-protein binding at the Indel24 site, and can also reveal allele-dependent NF-Y binding affinities.

Mutation of the CCAAT sites within the region suggests that Indel24 modulates NF-Y binding by altering the spacing between the flanking NF-Y binding sites, instead of donating an extra CCAAT binding site. The implication that the spacing between the CCAAT sites in this region is more critical for NF-Y binding is not unprecedented, as it is known from studies of the triple CCAAT CyclinB2 promoter that their precise alignments are required for their function in vivo (Bolognese 1999; Manni *et al*, 2001; Salsi *et al*, 2003). There is also ChIP-on-chip evidence to suggest that no CCAAT sites are closer than 24bp (Dolfini *et al*, 2009), and in vitro biochemical data of dual NF-Y binding to CCAAT boxes indicates that a distance of at least 24bp is required for them not to become mutually exclusive (Salsi *et al*, 2003; Liberati *et al*; 1999). This could explain why the extra CCAAT site on the Indel24 insertion element does not confer an incremental effect on transcriptional activity, as the spacing between S1-S2 and S2-S3 are 24bp and 23bp respectively (Figure 22A), and may not be conducive for an extra binding interaction. On the other hand, the 24bp insertion element increases the spacing between S1-S3 from 23bp to 47bp, which bears a closer resemblance to the 42bp spacing of S3-S4. This may be functionally more optimal, as it is known that distances between CCAAT motifs at NF-Y promoters are enriched at 32bp, 42bp and 53bp, corresponding to 3, 4 and 5 turns of the double helix, respectively (Dolfini *et al*, 2009).

The involvement of NF-Y is intriguing, as it is well-known to regulate the expression of genes involved in cell cycle control and progression (Muller *et al*, 2010). Both NF-Y and E2F1 have been linked to the development of multiple cancers including CRC (Dolfini *et al*, 2013; Morris *et al*, 2008), making *SHROOM2* a compelling candidate as a susceptibility gene. The functional role of *SHROOM2* is further investigated in Chapter 8.

In summary, the data presented here provides functional mechanistic evidence to support the eQTL effect of the Xp22.2 CRC risk locus, whereby a novel indel polymorphism can modulate NF-Y binding at the *SHROOM2* distal promoter region and appears to be the causal variant. The evidence reported here supports a growing number of studies, which highlights the value of functionally characterising disease-associated common genetic variation in the discovery of novel candidate susceptibility genes for complex traits (Moffat *et al*, 2007; Meyer *et al*, 2008; Musunuru *et al*, 2010; Harismendy *et al*, 2011; Nguyen *et al*, 2012).

Chapter 8

SHROOM2 as a candidate susceptibility gene for colorectal cancer

8.1 Introduction

The association between a colorectal cancer risk locus and *SHROOM2* expression suggests that *SHROOM2* may play a role in the predisposition to colorectal cancer. According to published literature, the SHROOM family of proteins are regulators of epithelial morphogenesis, characterized by their ability to bind F-actin and organise actomyosin networks (Dietz *et al*, 2006). SHROOM2, previously known as APXL, has a PDZ domain, a common structural domain of 80-90 amino acids found in signalling proteins. PDZ domain-containing proteins regulate diverse cellular processes and many signal transduction pathways (as reviewed by Subbaiah *et al*, 2011). More specifically, SHROOM2 has been shown to play a role in cell morphogenesis during endothelial and epithelial tissue development (Lee *et al*, 2009; Farber *et al*, 2011), cytoskeletal organisation (Dietz *et al*, 2006), tight-junction stabilisation (Etournay *et al*, 2007) and cell contractility and migration (Farber *et al*, 2011).

SHROOM2 was initially studied as a candidate gene in ocular albinism (Schiaffino *et al*., 1995), and has been shown to regulate melanosome biogenesis and localisation in the retinal pigment epithelium (Fairbank *et al*, 2006; Lee *et al*, 2009). This is intriguing, as abnormal retinal pigmentation, similar to the CHRPE lesions that are a component of the familial adenomatous polyposis syndrome, has previously been shown to be an extra-colonic feature of non-FAP CRC (Houlston *et al*, 1992; Dunlop *et al*, 1996)

Further suggesting a role in cancer, large-scale screens for mutations in somatic cancer have detected missense substitutions in the *SHROOM2* coding sequence in various tumours including colorectal cancers (Forbes *et al*, 2010; URL8.1). There is also evidence that *SHROOM2* is differentially expressed in medulloblastoma (Shou *et al*, 2015), and an intronic SNP within *SHROOM2* has been associated with genetic predisposition to prostate cancer (Eeles *et al*, 2013). *SHROOM2* has also recently been implicated in non-cancer disease traits; the *SHROOM2* gene was found to be associated with inherited predisposition to late-onset Alzheimer's disease (Meda *et al*, 2012), and the authors posited that it may be implicated in the formation of pathological tau proteins by mediating actin cytoskeletal changes.

SHROOM2's diverse cellular roles make it an interesting candidate gene for colorectal cancer development, although it has not been previously characterised in this respect. This chapter presents preliminary functional data that offers insight into role of SHROOM2 in colonic epithelial cells, by using a transient siRNA knockdown approach as well as tissue and subcellular localisation studies.

8.2 Methodological overview

8.2.1 Transcript-specific reverse-transcription PCR

PCR was performed on cDNA synthesised from human primary tissue and cell line RNA, as described in Chapter 2. Primers specific for the four different reported transcripts of *SHROOM2* were used (Table 8.1)

***SHROOM2* transcripts**	**Forward primer**	**Reverse primer**
T-001	GCCTCTTGGAAGGAACAG	GACACTGGGCATCTGCTTG
T-002	GCAGCCCTTGGTATGTG	GACACTGGGCATCTGCTTG
T-201	TGCGTGAGCTTGCCCATC	GACACTGGGCATCTGCTTG
T-003	CTGATCCAGCAAATGTGTGTAG	CAAAATAAATAGTGTCTCTTC

Table 8.1 Primer sequences used to specifically target and amplify the different transcripts of *SHROOM2.*

8.2.2 *SHROOM2* siRNA knockdown

siRNA knockdown of *SHROOM2* was performed as described in 7.2.5 using 10nM of siRNAs (Sigma) as detailed in Table 8.2. Transfected cells were incubated for 48 hours prior to seeding for phenotypic assays or RNA extraction for gene expression analysis.

Gene	**siRNA**	**Oligo ID**	**Sequence (5' - 3')**
SHROOM2	siRNA1	SASI_Hs01_00205221	GGUAUGUUCCCGAUAAGAA
SHROOM2	siRNA2	SASI_Hs02_00332240	CAAAGAGAAGACUGUGGAA
SHROOM2	siRNA3	SASI_Hs01_00205222	GAGACUUCUCCCAUAGCAA
Negative control	Scrambled	SIC001	

Table 8.2 The IDs and sequence of siRNAs (Sigma) used in this chapter.

8.2.3 Growth assays

Growth curves for colorectal cancer epithelial cell lines were performed after siRNA knockdown of *SHROOM2* for 48 hours. Cells were trypsinised, counted with a Coulter Counter (Beckman) and seeded at 2 x 10^6 in a T25 flasks for each time point. At each time point, cells were trypsinised and counted prior to RNA extraction. Doubling time was calculated from the cell counts using an online tool (Roth V, 2006) (URL8.2).

8.2.4 Scratch-wound assay

Scratch assays for colorectal cancer epithelial cell lines were performed after siRNA knockdown of *SHROOM2* for 48 hours, when the cell monolayer has achieved uniform confluence. For each well, a 200µL pipette tip was used to scratch a wound through the centre of the wells. The cells were washed with warm PBS gently to remove loose cells, and fresh low-serum media (1% FCS) was added to the wells. This is to reduce the effect of proliferation so that the effect of migration can be better observed. Cells were then placed in an Axiovert 200 live cell imaging system (Zeiss) and three fields of view were selected for each well. Time-lapse imaging was carried out every 15 minutes for 24 hours. The area of the wound was manually quantified and analysed using ImageJ.

8.2.5 *SHROOM2* siRNA knockdown in cell lines: whole-genome gene expression profiling and gene ontology analysis

SHROOM2 expression was knocked down with the three different siRNAs in five human cell lines from a variety of tissue types - DLD1 (CRC), SW480 (CRC), PNT (Prostate epithelium), HEK293 (embryonic kidney) and RPE1 (retinal epithelium). Knockdown experiments were replicated twice. After 48 hours, cell line RNA was extracted and knockdown confirmed with qRT-PCR. The RNA was then amplified and hybridised on the Illumina HumanHT-12 v4.0 Expression BeadChip Arrays (Illumina, USA), and gene expression data processed as described in 2.6.2. Differentially expressed genes that overlapped between all three siRNAs for each individual cell line were subjected to GO (gene ontology) terms enrichment analysis using the web-based tool GOrilla (Eran *et al*, 2009) (URL2.3). The analyses were performed using the running mode that compared the target list of genes to the background list of genes (n=14174) that were expressed and detected in the cell lines used for the knockdown experiment. Ontologies containing only a single gene were omitted.

8.2.6 Western blotting

Total protein and subcellular fractions were extracted as described in 2.5. Primary antibodies used are listed in Table 8.3.

Protein	Company	Catalogue no.	Type	Antibody dilution used
SHROOM2	Epitomics	#S0151	Rabbit polyclonal	1:500
β-actin	Sigma	#A1978	Mouse monoclonal	1:5000
HSP60	Sigma	#H3524	Mouse monoclonal	1:500
E-cadherin	Cell signalling	#3195	Rabbit monoclonal	1:1000
Lamin B1	Santa Cruz	#SC-6216	Goat polyclonal	1:500
Vimentin	Cell signalling	#5741	Rabbit monoclonal	1:500

Table 8.3 Details of the antibodies and dilutions used in this chapter.

SHROOM2 sera were produced by Dundee Cell Products using peptides[1] as listed in Table 8.4. Antibodies were affinity purified using cognate peptide bead columns.

Serum	Rabbit	Peptide sequence
CSA 71	71	CSAGAQEPPRASRAEKASQR
CSA 51	51	CSAGAQEPPRASRAEKASQR
AQA 71	71	AQAQPRGDRRPELTDRPWRSAH
AQA 51	51	AQAQPRGDRRPELTDRPWRSAH

Table 8.4 Details of the customised serum raised against sequences specific to *SHROOM2*.

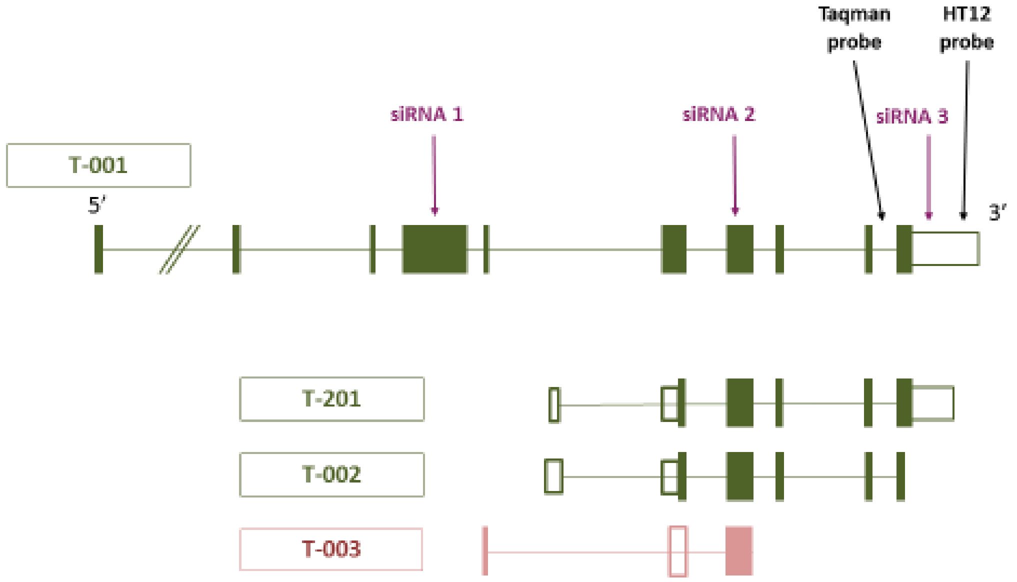
[1] Peptides were designed by Dr Susan Farrington, CCGG Human Genetics Unit

8.3 Results

8.3.1 Transcript-specific expression of *SHROOM2* in cell lines and colonic primary tissue

To begin studying the function of *SHROOM2* in colorectal epithelial cells, the *SHROOM2* gene and its four transcripts were first examined in a panel of cell lines and colonic tissue. This is important as the Illumina HT12 microarray probe detects only the largest transcript, whereas the Taqman Gene expression probe for qRT-PCR maps to all 3 protein-coding transcripts (Figure 8.1). To investigate which transcripts are expressed and relevant in colonic epithelial cells, RT-PCR using transcript-specific primers was performed on a panel of human CRC cell lines, normal colorectal mucosa (NM) and colorectal tumour samples. It appears that colorectal cancer cell lines and primary colorectal tissue (normal and tumour) predominantly express the canonical transcript *T-001* (Table 8.5 and 8.6), which is detected by both the Illumina HT-12 gene expression microarray and the qRT-PCR Taqman assay.

The transcript-specific primers were also tested against a panel of non-colorectal human cell lines (Table 8.7). This served as a positive control experiment showing that the primer sets were indeed working, and also demonstrated that there is tissue-specific expression of *SHROOM2* transcript isoforms. *SHROOM2* appears to be weakly expressed, if at all, in the lymphoblastoid and erythroleukemia cell lines, which is similar to what was previously observed in primary PBMCs. The retinal pigment epithelial cell line RPE1 appears to lack the canonical transcript and only expresses the short *T-201* transcript, whereas the prostate epithelial cell line PNT and the breast epithelial cell line MCF7 appears to express the shorter transcripts as well as the longer canonical transcript.



Name	Transcript ID	Length (bp)	Protein ID	Length (aa)	Biotype
T-001	ENST00000380913	7447	ENSP00000370299	1616	Protein coding
T-002	ENST00000452575	2276	ENSP00000406724	375	Protein coding
T-201	ENST00000418909	3597	ENSP00000415229	451	Protein coding
T-003	ENST00000493668	855	No protein product	-	Processed transcript

Figure 8.1 Summary of the *SHROOM2* gene (ENSG00000146950) on chromosome X: 9,754,496-9,917,483 and its 4 transcripts as detailed on Ensembl (GRCh37). Regions targeted by gene expression assays and siRNAs are also indicated.

Cell line	**Tissue of origin**	**rs5934683**	**T-001**	**T-002**	**T-201**	**T-003**
CACO2	CRC	TT	Present	Absent	Absent	Absent
COLO320	CRC	CC	Absent	Absent	Absent	Absent
DLD1	CRC	CC	Present	Absent	Absent	Absent
HCT116	CRC	TT	Present	Absent	Faint	Absent
HT29	CRC	CT	Absent	Absent	Absent	Absent
LOVO	CRC	TT	Present	Absent	Absent	Absent
RKO	CRC	CC	Present	Absent	Absent	Absent
SW48	CRC	CC	Absent	Absent	Absent	Absent
SW480	CRC	TT	Present	Absent	Absent	Absent
VACO425	CRC	CC	Present	Absent	Absent	Absent

Table 8.5 *SHROOM2* transcript-specific RT-PCR using RNA from CRC cell lines.

Patient	**Gender**	**rs5934683**	**Tissue**	**T-001**	**T-002**	**T-201**	**T-003**
CR77	F	CT	NM	Present	Absent	Absent	Absent
CR90	F	CT	NM	Present	Absent	Absent	Absent
			Tumour	Present	Absent	Absent	Absent
CR94	M	CC	NM	Present	Absent	Absent	Absent
			Tumour	Present	Absent	Absent	Absent
CR97	F	TT	NM	Present	Absent	Absent	Absent
CR102	M	CC	NM	Present	Absent	Absent	Absent
			Tumour	Present	Absent	Absent	Absent
CR104	M	CC	NM	Present	Absent	Absent	Absent
			Tumour	Present	Absent	Absent	Absent
CR142	M	TT	NM	Present	Absent	Absent	Absent
			Tumour	Present	Absent	Absent	Absent
CR152	M	TT	NM	Present	Absent	Absent	Absent
			Tumour	Present	Absent	Absent	Absent

Table 8.6 *SHROOM2* transcript-specific RT-PCR on normal colorectal mucosa and paired tumour tissue whenever available.

Cell line	**Tissue of origin**	**T-001**	**T-002**	**T-201**	**T-003**
HeLa	Cervical cancer	Present	Absent	Absent	Absent
MCF7	Breast cancer	Present	Faint	Absent	Absent
DU145	Prostate cancer	Present	Absent	Absent	Absent
PC3	Prostate cancer	Present	Absent	Absent	Absent
PNT	Prostate	Present	Present	Faint	Absent
K562	Erythroleukemia	Faint	Absent	Absent	Absent
CON A	Lymphoblastoid	Absent	Absent	Absent	Absent
CON C	Lymphoblastoid	Absent	Absent	Absent	Absent
RPE1	Retinal pigment epithelium	Absent	Absent	Present	Absent

Table 8.7 *SHROOM2* transcript-specific RT-PCR on non-CRC cell lines.

Next, the relative expression of *SHROOM2* in CRC and non-CRC cell lines was assessed by qRT-PCR (Figure 8.2). *SHROOM2* gene products were very lowly expressed in CRC cell lines COLO320, HT29, SW48 and the blood cell lines K562, ConA and ConC, which was consistent with the non-quantitative RT-PCR. The colorectal cancer cell lines CACO2, DLD1, HCT116 and SW480 were high expressors, and hence selected for siRNA knockdown and phenotypic functional studies. The genotypes of these cell lines did not appear to affect *SHROOM2* expression in a consistent manner, which is not surprising given the chromosomal and genomic instability that are inherent to these tumour cell lines.

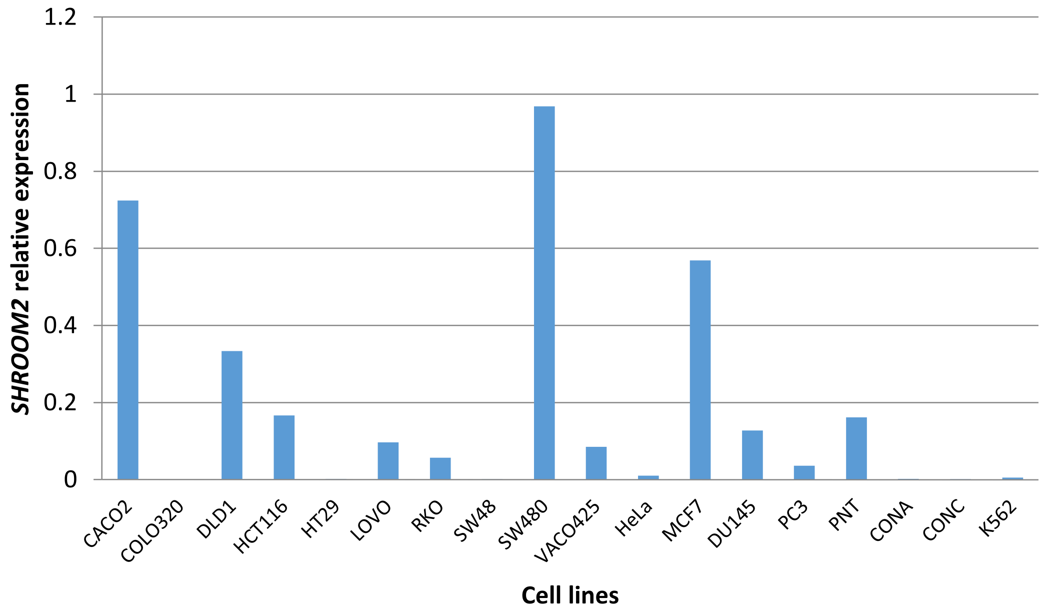


Figure 8.2 Relative expression of *SHROOM2* in a panel of colorectal cancer cell lines and non-colorectal cancer cell lines as quantified by qRT-PCR. *SHROOM2* was normalised to three reference genes *TBP*, *RPL30* and *EIF2B1*.

8.3.2 siRNA knockdown of *SHROOM2*

Having established that the *SHROOM2* canonical transcript (*T-001*) is the relevant transcript in colorectal epithelial tissue, transient knockdown of this transcript was performed in CRC cell lines using siRNA transfections. Two siRNAs were initially used; siRNA1 was chosen to target only the canonical transcript, whereas siRNA2 targets all four transcripts (Figure 8.1). HCT116 and DLD1 cell lines were initially chosen for knockdowns as they are adherent cell lines that transfect well and are good expressors of *SHROOM2*.

A transfection with a 48 hour incubation effectively knocked down *SHROOM2* at the mRNA and protein level, and the level of knockdown was similar when used individually or in combination (Figure 8.3 and 8.4). This suggests that the large transcript *SHROOM2-001* is the main, if not only, transcript expressed in these cell lines. It is also indicative that pooling of siRNAs did not appear to function synergistically to facilitate further degradation of *SHROOM2* mRNA.

The SHROOM2 antibody used for Western Blotting appears to detect multiple other protein bands of different sizes. However, they are unlikely to be SHROOM2 isoforms or SHROOM2 degradation products as their intensity does not change with knockdown using siRNA2, which targets all three protein-coding transcripts. Nevertheless, the non-specificity of this antibody renders it unsuitable for imaging and localisation studies. Efforts to obtain an antibody that is specific to SHROOM2 will be discussed in a separate section.

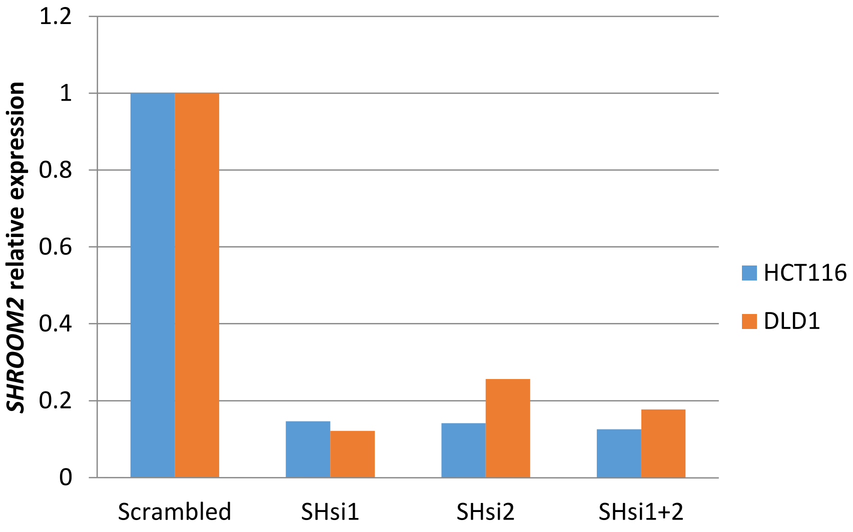


Figure 8.3 qRT-PCR showing *SHROOM2* expression normalised to *EIF2B1*, *RPL30* and *TBP*. *SHROOM2* was knocked down with the siRNAs individually and in combination for two CRC cell lines HCT116 and DLD1.

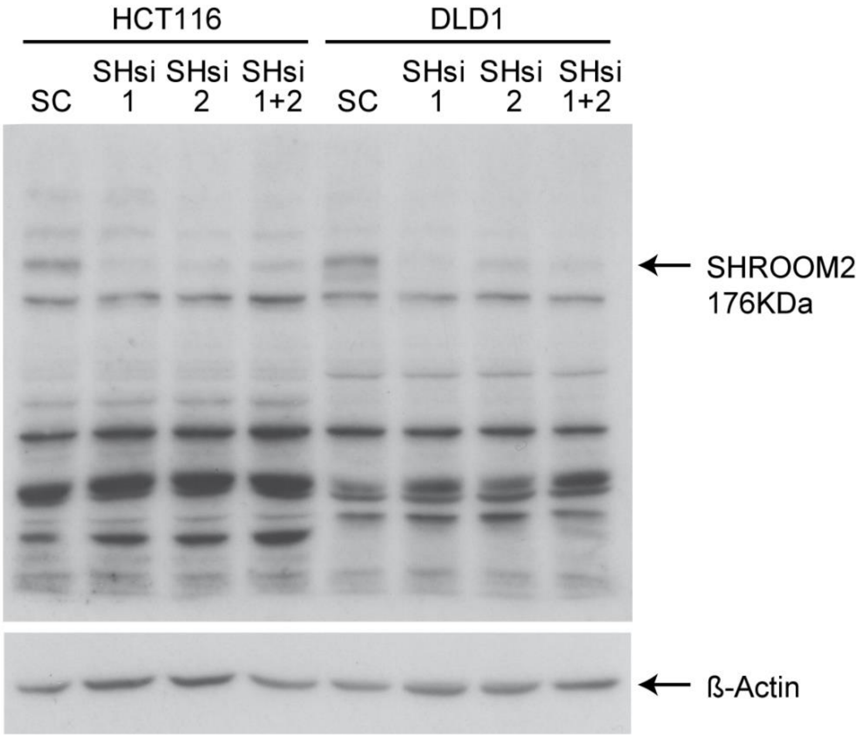


Figure 8.4 Western blots with a commercial SHROOM2 antibody. The band that was reduced with *SHROOM2* siRNA knockdown in both cell lines corresponds to the size of endogenous SHROOM2. β-actin was used as a loading control.

8.3.3 Transient knockdown of *SHROOM2* and growth assays in CRC cell lines

Growth, or increase in total cell number over time, is a good measure of a biological response because it is broadly defined and influenced by many different factors including mitogens, nutrient levels, changes in transport, membrane integrity, attachment factors and so forth. Cell numbers can be affected by death rate, mitotic rate, progression through the cell cycle, or even by changing the plateau density. Although not specific, growth curves can be used as a general screening tool to detect phenotypic changes when a gene product is suppressed.

To analyse the growth characteristics of colorectal epithelial cells when *SHROOM2* is knocked down, growth curves were established for 3 CRC cell lines with varying baseline levels of *SHROOM2* - SW480 is a high expressor, whereas DLD1 and HCT116 have moderate expression of the *SHROOM2* transcript. The cells were first transfected with siRNA for 48 hours and then trypsinised, counted and seeded in fresh media for the growth curves. The population doubling time for the cells were then calculated as a measure of cell proliferation. siRNA1 and siRNA2 are initially pooled to reduce the number of experiments performed. *SHROOM2* mRNA level was measured at each time point to ensure that the reduction in expression was maintained throughout the experiment, and Western Blotting performed at time point 0 to confirm depletion of SHROOM2 protein (Figure 8.5). Knock down of *SHROOM2* appeared effective with a reduction in mRNA expression of >85% at 0 hours. This effect diminished over time with a residual reduction of >50% at 72 hours, which is expected of transient siRNA knockdowns. It is interesting to note that *SHROOM2* mRNA expression for both the control and knocked down cells appeared to increase with the time in culture, which may reflect the confluence or density of cells. The high levels at 0 hours (matching that of 72 hours) make this more likely - although the number of cells at 0 hours was low, they would have been almost confluent prior to trypsin digest and seeding. As RNA was extracted immediately after trypsin digest, the levels of cellular mRNA at this time point would be representative of that of a confluent phase. This was consistently observed in all the cell lines. This apparent relationship between *SHROOM2* expression and cell confluence may relate to, and provide support to its reported role in tight junction stabilisation. (Etournay *et al*, 2007). This does not, however, exclude confounders such as growth factors and nutrient composition in the growth media, which can also change as a function of time.

When the growth curves were plotted for the replicates of each cell line experiment, the rate of growth for DLD1 appeared to be slower when *SHROOM2* was transiently knocked down. (Figure 8.6). This was not seen with the other cell lines. Mann-Whitney U test on the population doubling time confirmed that there was a significant difference in doubling time (*p*-value=0.01) between DLD1 cells treated with scrambled siRNA and those treated with the *SHROOM2* siRNAs. The knockdown cells had a 6.6 hour increase in average doubling time. To rule out off-target effects, the experiment was repeated with the siRNAs singularly to ensure that this effect is specific to *SHROOM2*. A third siRNA that targets the 3’ end of *SHROOM2-001* was used (Figure 8.1) for further confidence. All three siRNAs appeared to knockdown *SHROOM2* mRNA levels to similar degrees (~80%), with siRNA1 maintaining the knockdown most effectively (Figure 8.7). The DLD1 growth curves for the individual siRNAs revealed that the slowing of growth rates was only present with siRNA2 and not replicated with siRNA1 and siRNA3 (Figure 8.8), indicating that this is more likely to be an off-target effect of siRNA2 rather that an effect attributable to *SHROOM2* specific RNA degradation.

In summary, transient depletion of *SHROOM2* did not appear to affect the population doubling time of CRC cell lines DLD1, HCT116 and SW480. The slowing of growth observed in DLD1 when *SHROOM2* was knocked down is likely to be due to siRNA off targeting, as this effect was not replicable using siRNAs with comparable gene silencing efficiencies.

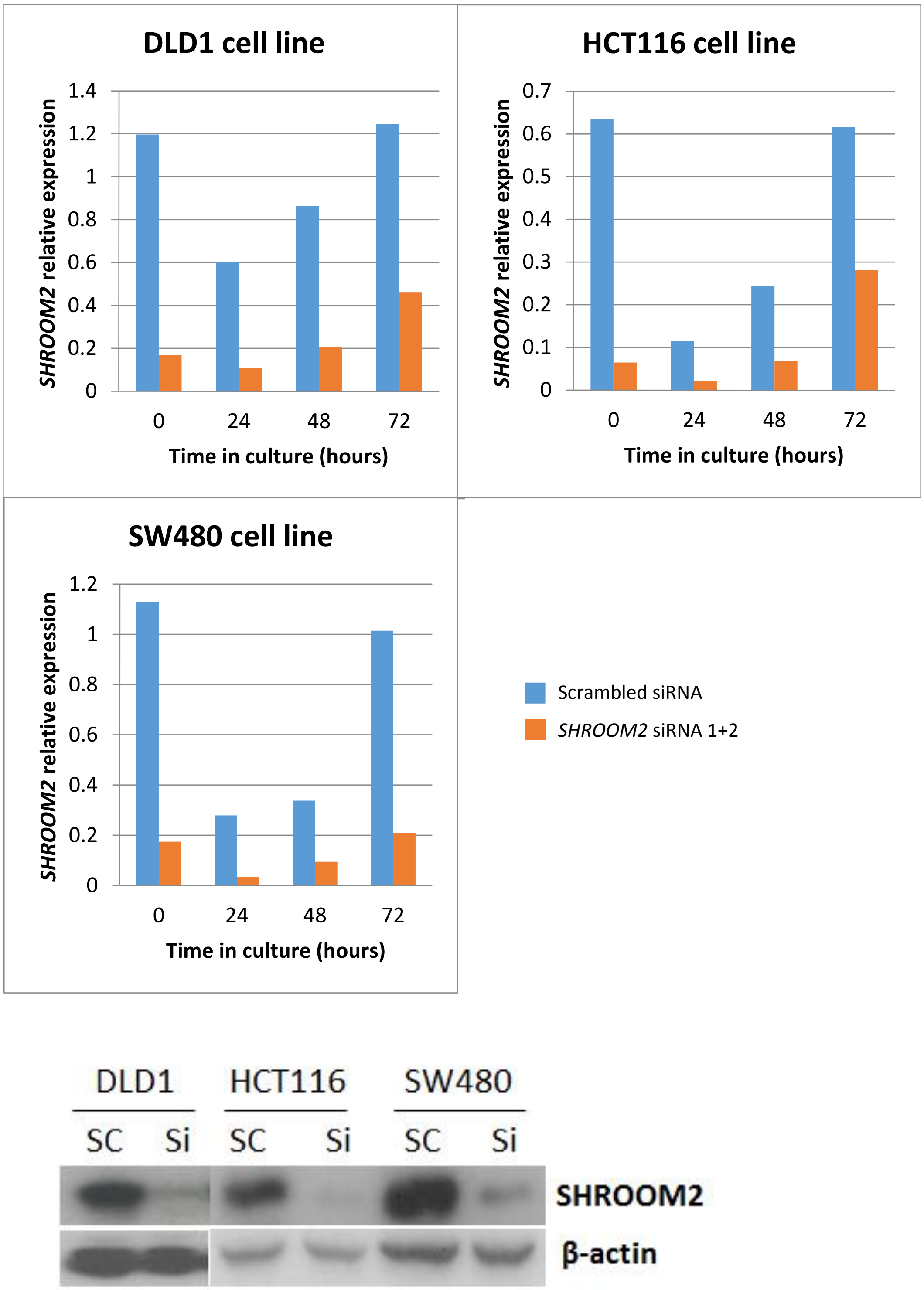
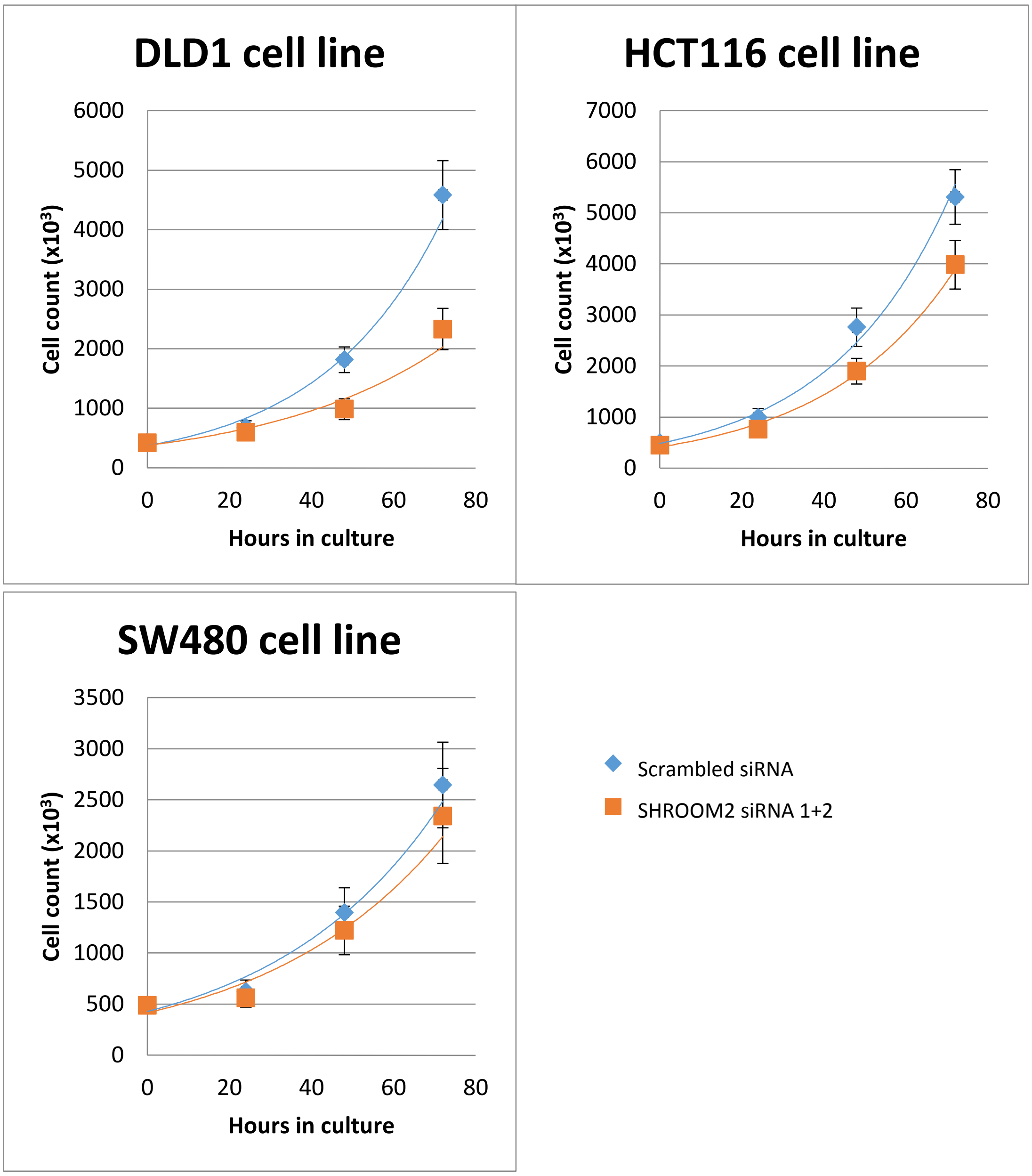


Figure 8.5 qRT-PCR of CRC cell lines when *SHROOM2* was knocked down with siRNA1+2. *SHROOM2* normalised to *ACTB*. Western blots of total protein extracted at time point = 0. Graphs and blots shown are representative of replicates. SC=scrambled control, Si=*SHROOM2* siRNA 1+2.



Cell line	Average doubling time (hours)		*p*-value
	Scrambled	*SHROOM2* siRNA 1+2	
DLD1	20.9	27.5	0.01
HCT116	19.5	21.5	0.43
SW480	26.4	27.3	1

Figure 8.6 Growth curves of colorectal cancer cell lines when *SHROOM2* was knocked down with siRNA1+2. Experiments replicated at least 4 times, error bars=SEM. *p*-values are reported for the Mann-Whitney U test.

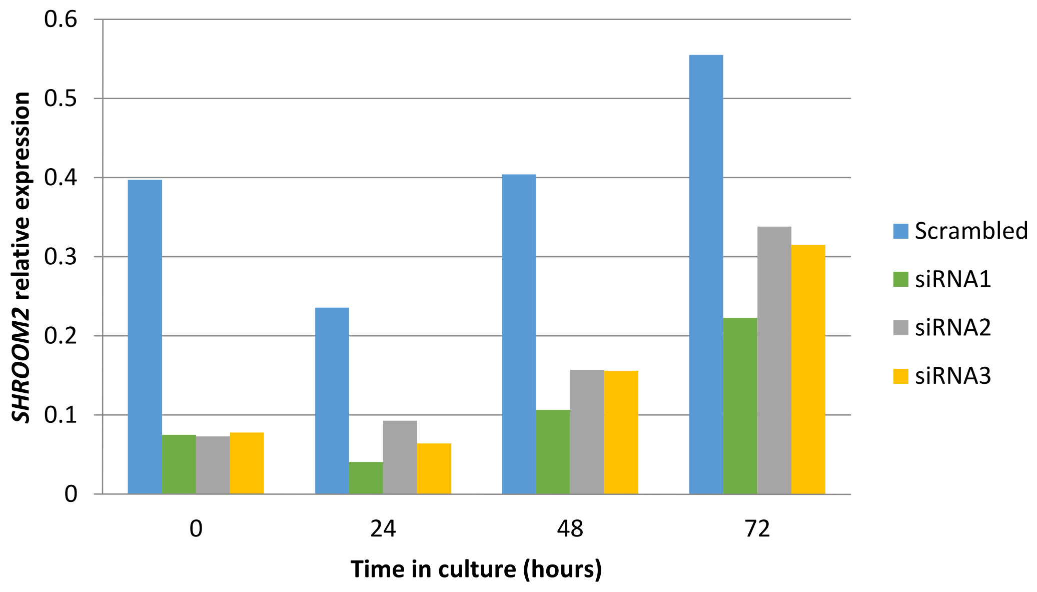


Figure 8.7 qRT PCR of DLD1 with *SHROOM2* knocked down with three different siRNAs singularly. *SHROOM2* normalised to *ACTB*.

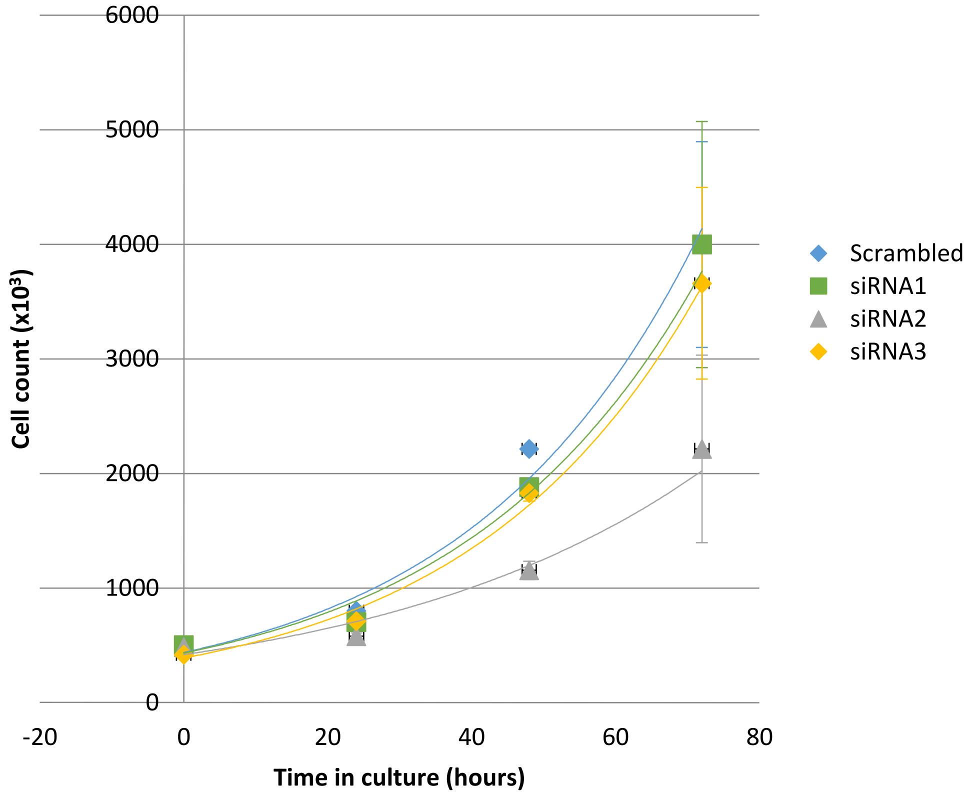


Figure 8.8 Growth curves of DLD-1 when *SHROOM2* was knocked down with three different siRNAs singularly. Experiments replicated 2 times, error bars=SEM.

8.3.4 Transient knockdown of *SHROOM2* and scratch-wound assays in CRC cell lines

The in vitro scratch-wound assay is a straightforward, reproducible assay commonly used to measure basic cell migration parameters. Creation of a “scratch” gap in the confluent cell monolayer induces cells on the edge of the gap to polarise and migrate toward the opening to close the “scratch” until new cell-cell contacts are re-established. It mimics to some extent migration of cells in vivo, and can be useful to study the regulation of cell migration by cell-cell interactions.

As *SHROOM2* appears to regulate endothelial sprouting, migration and angiogenesis (Farber 2011), it was hypothesised that *SHROOM2* may also play a role in colonic epithelial cell migration. To test this hypothesis, in vitro scratch assays using time-lapse imaging were performed after siRNA knockdown of *SHROOM2* in DLD1 cell lines. The closure of wound gap was quantified by measuring the remaining area of the gap at multiple time points. The depletion of *SHROOM2* did not appear to affect the rate of wound closure in DLD1 (Figure 8.9A). This experiment was repeated in SW480 and CACO2 cell lines and showed inconsistent results. In SW480, *SHROOM2* knockdown slowed wound closure, whereas this had the opposite effect in CACO2 (Figure 8.9B). *SHROOM2* expression levels were quantified with qRT-PCR after the final time-point and the knockdown of *SHROOM2* was >80% for DLD1 and SW480 but was less so at 50% for CACO2 (Figure 8.10). This experiment has been performed only once in the SW480 and CACO2 cell lines, and will require technical replication.

In summary, within the remit and limitations of this simple scratch-wound assay, *SHROOM2* did not appear to have a consistent effect on the rate of wound closure in a monolayer of CRC cells.

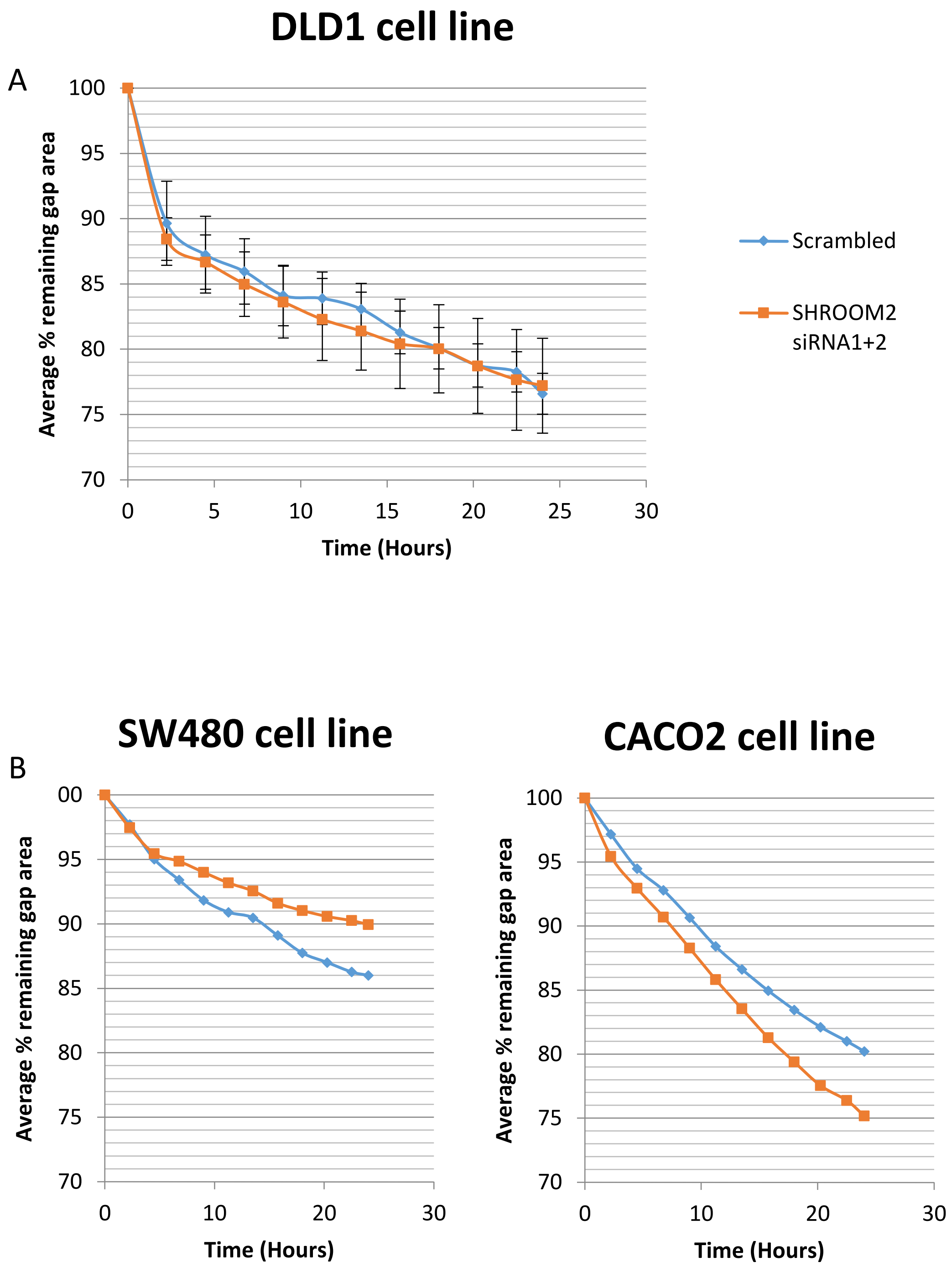


Figure 8.9 A) Scratch wound assay - the remaining gap area over a 24 hour time course was quantified as a measure of wound closure in DLD1 cell line monolayer. Transient knockdown of SHROOM2 with siRNA was performed prior to introducing the scratch wound. Experiments replicated 4 times, error bars=SEM. B) Scratch wound assay similarly performed on two other CRC cell lines (SW480 and CACO2).

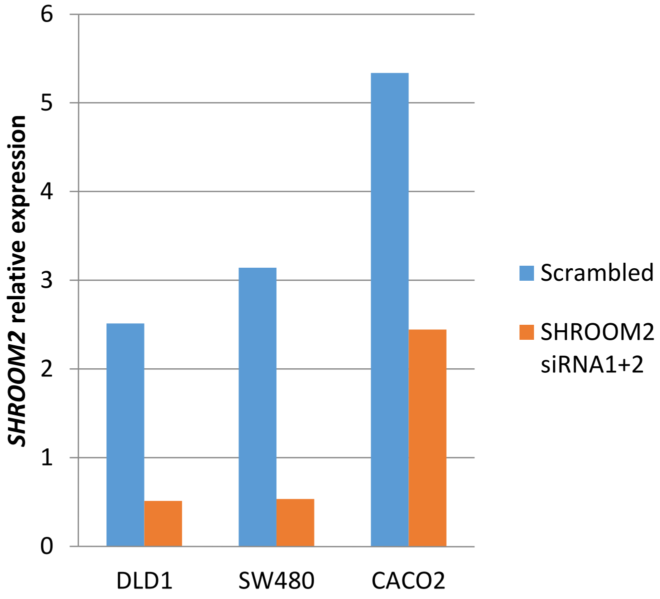


Figure 8.10 qRTPCR showing depletion of *SHROOM2* at the end of the scratch wound time course (24hours) for the three cell lines used. *SHROOM2* normalised to *ACTB*. Results representative of the replicates are shown for DLD1.

8.3.5 Microarray gene expression analysis of normal mucosa and cell lines in relation to *SHROOM2* expression

To gain further insight into the function of *SHROOM2*, microarray data from whole-genome gene expression profiling of normal mucosa were first examined.

Of the 21937 genes detected in the normal mucosa, the expression levels of 4570 genes were correlated with *SHROOM2* expression (pers. comm. Grimes). 2390 of these were positively correlated, whereas 2180 were negatively correlated. Under the working assumption that functionally related genes are more likely to be co-expressed (Eisen *et al*, 1998; Hughes *et al*, 2000; Kim *et al*, 2001), these genes were subjected to gene ontology analysis using GOrilla. The top twenty most significant GO terms for all correlated genes, positively correlated genes and negatively correlated genes are presented in Table 8.8 – 8.10. There is a striking presence of GO terms implicating the cell cycle when all correlated genes are analysed together. The positively correlated genes appear to be enriched for metabolic/catabolic processes, whereas the negatively correlated genes were predominantly enriched for cell cycle processes followed by metabolic processes.

Description	*p*-value	FDR *q*-value	Enrichment factor	No. of genes
cell cycle process	3.80E-13	4.84E-09	1.82	176
mitotic cell cycle process	3.01E-12	1.92E-08	1.88	132
cell cycle	9.51E-10	4.04E-06	1.81	117
mitotic cell cycle	6.19E-09	1.97E-05	1.96	86
cell cycle G1/S phase transition	4.72E-08	1.20E-04	2.6	40
G1/S transition of mitotic cell cycle	4.72E-08	1.00E-04	2.6	40
cell cycle phase transition	1.83E-07	3.33E-04	2.08	60
mitotic cell cycle phase transition	2.52E-07	4.02E-04	2.07	59
regulation of mitotic cell cycle	3.25E-07	4.60E-04	2.05	62
regulation of cell cycle process	8.06E-07	1.03E-03	1.93	68
DNA strand elongation involved in DNA replication	8.61E-07	9.97E-04	4.43	16
regulation of cell cycle	1.02E-06	1.09E-03	1.67	106
DNA strand elongation	1.46E-06	1.43E-03	4.3	16
rRNA metabolic process	2.02E-06	1.84E-03	2.39	36
chromosome organization	4.40E-06	3.73E-03	2	52
regulation of catalytic activity	5.24E-06	4.17E-03	1.84	69
regulation of transferase activity	5.49E-06	4.11E-03	2.28	42
cell division	5.70E-06	4.03E-03	1.86	63
rRNA processing	5.76E-06	3.86E-03	2.44	35
DNA metabolic process	8.41E-06	5.36E-03	1.56	113

Table 8.8 Gene ontology analysis of the genes that are correlated in expression to *SHROOM2* in the normal mucosa.

Description	*p*-value	FDR *q*-value	Enrichment factor	No. of genes
lipid metabolic process	1.94E-10	2.47E-06	1.54	189
fatty acid oxidation	8.72E-09	5.56E-05	3.37	25
cellular lipid metabolic process	1.09E-08	4.62E-05	1.56	149
lipid oxidation	1.36E-08	4.33E-05	3.31	25
enzyme linked receptor protein signaling pathway	2.79E-07	7.11E-04	1.49	147
cellular lipid catabolic process	6.45E-07	1.37E-03	2.29	37
fatty acid beta-oxidation	6.49E-07	1.18E-03	3.33	19
lipid modification	1.87E-06	2.98E-03	2.23	36
phosphate-containing compound metabolic process	4.49E-06	6.36E-03	1.29	259
fatty acid catabolic process	5.42E-06	6.90E-03	2.72	22
lipid catabolic process	6.15E-06	7.12E-03	1.89	49
phosphorus metabolic process	6.76E-06	7.18E-03	1.28	263
monocarboxylic acid catabolic process	8.48E-06	8.31E-03	2.48	25
carnitine transport	1.07E-05	9.74E-03	5.49	8
amino-acid betaine transport	1.07E-05	9.09E-03	5.49	8
fatty acid metabolic process	1.36E-05	1.08E-02	1.75	57
regulation of plasma membrane organization	2.01E-05	1.51E-02	2.6	21
ammonium ion metabolic process	2.31E-05	1.63E-02	2.01	36
organic hydroxy compound metabolic process	3.22E-05	2.16E-02	1.54	83
vacuolar transport	4.19E-05	2.67E-02	2.38	23

Table 8.9 Gene ontology analysis of the genes that are positively correlated in expression to *SHROOM2* in the normal mucosa.

Description	*p*-value	FDR *q*-value	Enrichment factor	No. of genes
cell cycle process	5.09E-35	6.49E-31	2.22	236
mitotic cell cycle process	1.50E-32	9.56E-29	2.41	186
mitotic cell cycle	1.53E-31	6.50E-28	2.82	137
cell cycle	4.92E-26	1.57E-22	2.3	164
heterocycle metabolic process	7.92E-23	2.02E-19	1.38	629
nucleobase-containing compound metabolic process	1.36E-22	2.89E-19	1.39	607
cellular aromatic compound metabolic process	3.10E-22	5.64E-19	1.38	627
cellular macromolecule metabolic process	1.77E-21	2.82E-18	1.28	835
cellular nitrogen compound metabolic process	2.76E-21	3.91E-18	1.35	656
ncRNA metabolic process	2.99E-21	3.81E-18	2.51	112
gene expression	8.57E-21	9.92E-18	1.98	182
DNA metabolic process	1.96E-20	2.08E-17	2.05	164
organic cyclic compound metabolic process	4.95E-20	4.85E-17	1.34	643
nucleic acid metabolic process	1.03E-19	9.38E-17	1.39	548
nitrogen compound metabolic process	4.13E-19	3.50E-16	1.31	692
macromolecule metabolic process	1.11E-18	8.80E-16	1.24	895
primary metabolic process	3.20E-18	2.39E-15	1.2	1020
chromosome organization	5.16E-18	3.65E-15	2.76	79
cellular component organization or biogenesis	4.05E-17	2.71E-14	1.31	627
cell division	5.39E-16	3.43E-13	2.4	90

Table 8.10 Gene ontology analysis of the genes that are negatively correlated in expression to *SHROOM2* in the normal mucosa.

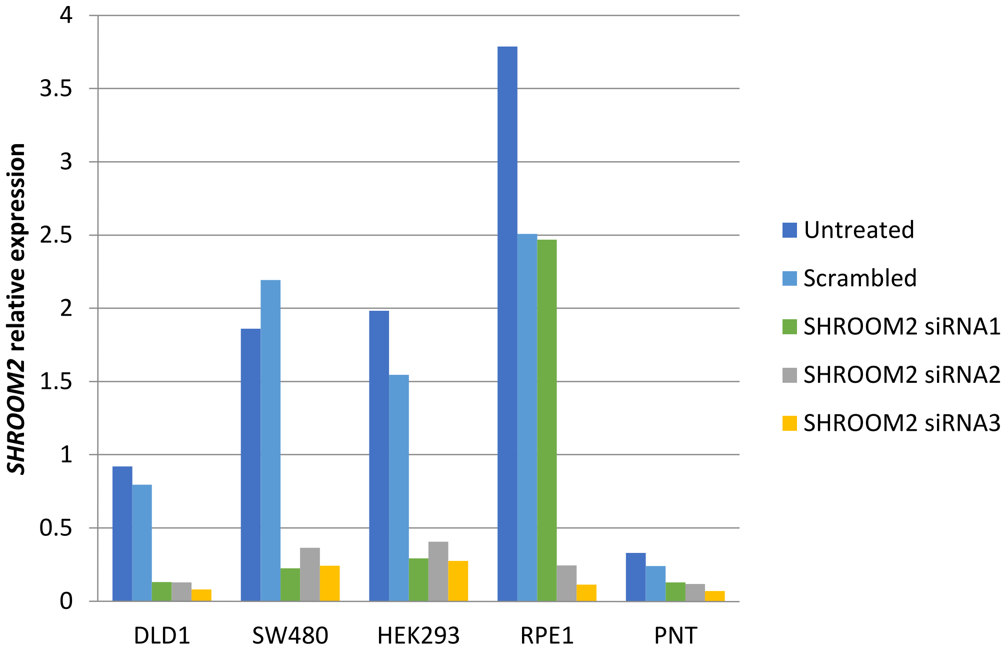
To evaluate further the role of *SHROOM2* in cancer and non-cancer cells, siRNA knockdown of *SHROOM2* was performed in two CRC cell lines (DLD1 and SW480) and three non-cancer cell lines (HEK293, RPE1 and PNT), using three different siRNAs (Figure 8.13A). siRNA1 was not effective in RPE1; this was expected given that RPE1 expresses only the shorter transcript *SHROOM2-201* which lacks the target exon of siRNA1, as shown in Table 8.8 and Figure 8.1.

Gene expression microarray analysis showed that there were 14174 unique genes detected, and the number of genes with differential expression varied between cell lines (Figure 8.13B). Gene ontology analysis was performed for genes that are downregulated and upregulated upon depletion of *SHROOM2*, for each cell line individually (Table 8. - 8.). Overall, cell-cycle and cell-division related genes appear to be overrepresented, most notably within the downregulated genes in DLD1, HEK293 and PNT cell lines. This association with cell-cycle genes is consistent with the normal mucosa GO analysis, strongly suggesting *SHROOM2* has a regulatory function of the cell cycle. Interestingly, *SHROOM2* has been reported to be a centrosome-associated protein in a mouse endothelial cell line, and plays a role in the regulation of centrosome duplication (Farber, 2012). Moreover, Xenopus *SHROOM2* has been shown to regulate gamma tubulin (Fairbank *et al*, 2006), which is found primarily in centrosomes and spindle pole bodies. Though speculative, the inference that *SHROOM2* has a similar role in human colorectal epithelial cells is an attractive one, as centrosome defects are known to promote chromosomal instability (Ganem *et al*, 2009), a common and important pathway in the aetiology of colorectal cancer.

In the RPE1 cell line, the genes upregulated with knockdown of *SHROOM2* are enriched with protein localisation and transport genes, which would fit it with *SHROOM2*'s known regulatory function in melanosome biogenesis and localisation in the retinal pigment epithelium (Fairbank 2006). In the non-cancer HEK293, RPE1 and PNT cell lines, the GO analysis also points towards genes with a function in cell polarity, morphogenesis and organelle organisation, which could relate to *SHROOM2*'s interaction with the actin-cytoskeleton and cell-junction proteins (Etournay *et al*, 2007). In the DLD1 cell line only, there is enrichment of *TGF-β/SMAD* signalling genes in the upregulated genes, which is intriguing given the involvement of this pathway in the pathogenesis of highly penetrable colorectal cancer mutational syndromes and colonic crypt homeostasis (Hardwick *et al*, 2008; Bellam *et al*, 2010; Reynolds *et al*, 2014). However, this should be interpreted with caution as a number of these enriched processes do not survive multiple testing correction.

The microarray gene expression analysis of these siRNA knockdown experiments has also highlighted the well-recognised caveat of the immunostimulatory "side effects" of siRNA treatments. There is a component of immune response, viral response and cell death genes observed in varying degrees across all cell lines, consistent with the knowledge that these effects are cell-type specific. In line with a stimulatory effect, these tend to be enriched for within the upregulated genes, although in RPE1 there was markedly strong enrichment within the downregulated genes. This idiosyncrasy could reflect a stronger immunostimulatory effect of the scrambled control siRNA in RPE1 cells. Hence, caution should be exercised when interpreting the results of these experiments, and a reduction in the concentration of siRNA used should be considered for future experiments, in particular with RPE1 cell line.

A)



B)

Cell line	Cell type	No. of downregulated genes	No. of upregulated genes
DLD1	Colorectal cancer	792	1015
SW480	Colorectal cancer	1353	660
HEK293	Human embryonic kidney	593	759
RPE1	Retinal pigment epithelium	1518	1542
PNT	Prostate	861	895

Figure 8.11 A) qRTPCR of cell lines selected for transcriptomic analysis showing siRNA knockdown of *SHROOM2*. *SHROOM2* normalised to *RPL30*. Graphs are representative of both replicates. B) Number of genes changed with siRNA knockdown of *SHROOM2*.

DOWNREGULATED GENES IN DLD1 UPON DEPLETION OF *SHROOM2*

Description	*p*-value	FDR *q*-value	Enrichment factor	No. of genes
cell cycle	3.79E-08	4.79E-04	2.15	57
mitotic cell cycle	3.52E-07	2.23E-03	2.31	42
modification-dependent protein catabolic process	4.33E-06	1.82E-02	2.32	34
cellular macromolecule catabolic process	5.78E-06	1.83E-02	1.88	54
ubiquitin-dependent protein catabolic process	6.27E-06	1.59E-02	2.32	33
modification-dependent macromolecule catabolic process	6.42E-06	1.35E-02	2.28	34
proteasome-mediated ubiquitin-dependent protein catabolic process	1.34E-05	2.43E-02	2.47	27
proteasomal protein catabolic process	2.23E-05	3.53E-02	2.41	27
proteolysis involved in cellular protein catabolic process	2.32E-05	3.26E-02	2.15	34

UPREGULATED GENES IN DLD1 UPON DEPLETION OF *SHROOM2*

Description	*p*-value	FDR *q*-value	Enrichment factor	No. of genes
negative regulation of response to stimulus	1.58E-04	>0.05	1.48	85
enzyme linked receptor protein signaling pathway	2.26E-04	>0.05	1.56	65
transmembrane receptor protein serine/threonine kinase signaling pathway	2.27E-04	>0.05	2.3	22
SMAD protein complex assembly	3.13E-04	>0.05	10.07	4
transforming growth factor beta receptor signaling pathway	3.14E-04	>0.05	2.56	17

Table 8.11 Gene ontology analysis of genes that are differentially expressed with depletion of *SHROOM2* in DLD1 cells. Where FDR *q*-values are non-significant, the top 5 highest ranked enriched processes are presented.

DOWNREGULATED GENES IN SW480 UPON DEPLETION OF *SHROOM2*

Description	*p*-value	FDR *q*-value	Enrichment factor	No. of genes
liver development	3.11E-04	>0.05	2.72	15
mature ribosome assembly	4.08E-04	>0.05	8.94	4
cardiac septum morphogenesis	6.77E-04	>0.05	3.27	10
ER-associated ubiquitin-dependent protein catabolic process	8.30E-04	>0.05	3.19	10
protein folding in endoplasmic reticulum	8.96E-04	>0.05	7.66	4

UPREGULATED GENES IN SW480 UPON DEPLETION OF *SHROOM2*

Description	*p*-value	FDR *q*-value	Enrichment factor	No. of genes
mitochondrial ATP synthesis coupled proton transport	2.80E-04	>0.05	8.02	5
hydrogen ion transmembrane transport	3.28E-04	>0.05	3.45	11
positive regulation of molecular function	4.83E-04	>0.05	1.51	66
transmembrane receptor protein tyrosine kinase signaling pathway	5.94E-04	>0.05	1.76	37
ribose phosphate metabolic process	6.00E-04	>0.05	2.45	17

Table 8.12 Gene ontology analysis of genes that are differentially expressed with depletion of *SHROOM2* in SW480 cells. Where FDR *q*-values are non-significant (>0.05), the top 5 highest ranked enriched processes are presented.

DOWNREGULATED GENES IN HEK293 UPON DEPLETION OF *SHROOM2*

Description	*p*-value	FDR *q*-value	Enrichment factor	No. of genes
mitotic cell cycle	1.05E-06	1.33E-02	2.57	32
cell division	2.28E-06	1.44E-02	2.73	27
chromosome organization involved in meiosis	5.14E-06	2.16E-02	9.57	7
mitotic nuclear division	1.34E-05	4.24E-02	2.98	20
cell cycle process	1.37E-05	3.47E-02	1.86	51
cell cycle	1.51E-05	3.19E-02	2.08	38
nuclear division	1.88E-05	3.39E-02	2.74	22

UPREGULATED GENES IN HEK293 UPON DEPLETION OF *SHROOM2*

Description	*p*-value	FDR *q*-value	Enrichment factor	No. of genes
protein modification by small protein conjugation or removal	1.47E-04	>0.05	1.73	48
Golgi vesicle transport	1.80E-04	>0.05	2.34	22
negative regulation of cell communication	1.86E-04	>0.05	1.56	66
regulation of cellular respiration	1.92E-04	>0.05	8.36	5
establishment of cell polarity	2.08E-04	>0.05	4.25	9

Table 8.13 Gene ontology analysis of genes that are differentially expressed with depletion of *SHROOM2* in HEK293 cells. Where FDR *q*-values are non-significant (>0.05), the top 5 highest ranked enriched processes are presented.

DOWNREGULATED GENES IN RPE1 UPON DEPLETION OF *SHROOM2*

Description	*p*-value	FDR *q*-value	Enrichment factor	No. of genes
type I interferon signaling pathway	8.91E-14	1.13E-09	5.11	27
response to virus	2.27E-10	1.43E-06	2.64	48
defense response to virus	2.20E-09	9.29E-06	3.03	34
negative regulation of viral process	6.46E-08	2.04E-04	3.45	23
negative regulation of viral life cycle	2.28E-07	5.76E-04	3.34	22
cytokine-mediated signaling pathway	5.03E-07	1.06E-03	1.95	58
negative regulation of viral genome replication	1.27E-06	2.29E-03	4.05	15
regulation of viral genome replication	1.82E-06	2.87E-03	3.31	19
response to other organism	2.07E-06	2.90E-03	1.88	57
response to biotic stimulus	3.60E-06	4.55E-03	1.68	77
negative regulation of multi-organism process	3.87E-06	4.45E-03	2.56	27
regulation of viral process	4.27E-06	4.50E-03	2.3	33
response to external biotic stimulus	4.52E-06	4.39E-03	1.69	74
cellular macromolecule metabolic process	7.44E-06	6.71E-03	1.15	576
regulation of viral life cycle	7.60E-06	6.41E-03	2.27	32
defense response to other organism	1.07E-05	8.47E-03	2.01	41
interferon-gamma-mediated signaling pathway	2.05E-05	1.53E-02	2.94	18
regulation of symbiosis, encompassing mutualism through parasitism	2.52E-05	1.77E-02	2.12	33
response to interferon-alpha	3.46E-05	2.30E-02	5.4	8
regulation of multi-organism process	3.57E-05	2.25E-02	1.83	47

UPREGULATED GENES IN RPE1 UPON DEPLETION OF *SHROOM2*

Description	*p*-value	FDR *q*-value	Enrichment factor	No. of genes
multi-organism cellular process	1.45E-09	1.83E-05	1.87	94
viral process	2.16E-09	1.36E-05	1.86	93
symbiosis, encompassing mutualism through parasitism	2.16E-09	9.09E-06	1.86	93
interspecies interaction between organisms	4.40E-09	1.39E-05	1.78	102
establishment of protein localization	1.44E-08	3.64E-05	1.55	154
establishment of protein localization to membrane	1.77E-08	3.72E-05	2.37	47
translation	8.03E-08	1.45E-04	2.14	53
establishment of localization in cell	1.23E-07	1.94E-04	1.43	187
single-organism intracellular transport	1.36E-07	1.91E-04	1.57	128
intracellular transport	1.41E-07	1.78E-04	1.5	152
protein transport	1.44E-07	1.66E-04	1.52	143
translational elongation	1.97E-07	2.08E-04	2.4	39
Golgi vesicle transport	3.76E-07	3.66E-04	2.29	41
nuclear-transcribed mRNA catabolic process	4.34E-07	3.91E-04	2.39	37
nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	4.93E-07	4.15E-04	2.77	28
organic substance transport	6.03E-07	4.77E-04	1.39	192
protein targeting to ER	6.60E-07	4.91E-04	2.79	27
protein localization to endoplasmic reticulum	7.18E-07	5.04E-04	2.72	28
cellular component organization or biogenesis	7.48E-07	4.97E-04	1.22	420
intracellular protein transport	8.72E-07	5.51E-04	1.65	92

Table 8.14 Gene ontology analysis of genes that are differentially expressed with depletion of *SHROOM2* in RPE1 cells. The top 20 highest ranked enriched processes are presented.

DOWNREGULATED GENES IN PNT UPON DEPLETION OF *SHROOM2*

Description	*p*-value	FDR *q*-value	Enrichment factor	No. of genes
organelle assembly	6.20E-09	7.83E-05	2.83	38
microtubule-based process	3.35E-08	2.12E-04	2.37	47
cilium organization	3.62E-08	1.52E-04	3.88	22
cell division	1.25E-07	3.95E-04	2.46	40
cilium assembly	2.58E-07	6.52E-04	3.76	20
organelle organization	6.70E-07	1.41E-03	1.45	151
single-organism organelle organization	8.16E-07	1.47E-03	1.55	115
mitotic cell cycle process	8.29E-07	1.31E-03	1.9	62
cellular component assembly involved in morphogenesis	1.98E-06	2.79E-03	2.94	24
cell cycle process	3.05E-05	3.86E-02	1.61	73
intraciliary transport	3.65E-05	4.20E-02	5.91	8

UPREGULATED GENES IN PNT UPON DEPLETION OF *SHROOM2*

Description	*p*-value	FDR *q*-value	Enrichment factor	No. of genes
regulation of biological quality	4.43E-06	5.60E-02	1.4	154
cellular component organization or biogenesis	3.78E-05	>0.05	1.24	247
interspecies interaction between organisms	3.80E-05	>0.05	1.73	57
enzyme linked receptor protein signaling pathway	5.02E-05	>0.05	1.67	62
cellular component organization	6.05E-05	>0.05	1.23	244

Table 8.15 Gene ontology analysis of genes that are differentially expressed with depletion of *SHROOM2* in PNT cells. Where FDR *q*-values are non-significant (>0.05), the top 5 highest ranked enriched processes are presented.

8.3.6 SHROOM2 protein localisation and function

The appropriate localisation of a protein is fundamental as it provides the physiological context for their function. *SHROOM2* has been showed from the gene expression studies in Chapter 6 to be expressed in the colonic normal mucosa. However, this is a highly heterogeneous tissue that consists of epithelium, connective tissue (lamina propria) and a thin muscle layer (muscularis mucosae) (Figure 8.19). They are morphologically distinct yet functionally interdependent, for example, the mesenchymal cells of the lamina propria orchestrate the microenvironment of the epithelial cells and regulate the stem cell niche within the crypts. The epithelium is further divided into more subtypes (e.g. enterocytes and goblet cells), and it is well-recognised that there is a differential distribution of gene expression along the colonic crypt-lumen axis as well as a proliferative and differentiation hierarchy. In the context of cancer, the spatial distribution of this gene product is highly relevant, as disruptions in the crypt dynamics and homeostasis is one of the early steps in malignant transformation of the colonic epithelium. On a cellular level, the eukaryotic cell is organised into membrane-covered compartments that are characterised by specific sets of proteins and biochemically distinct cellular processes. Hence, the appropriate tissue distribution as well as subcellular localisation of endogenous SHROOM2 would provide key insights into its functional role.

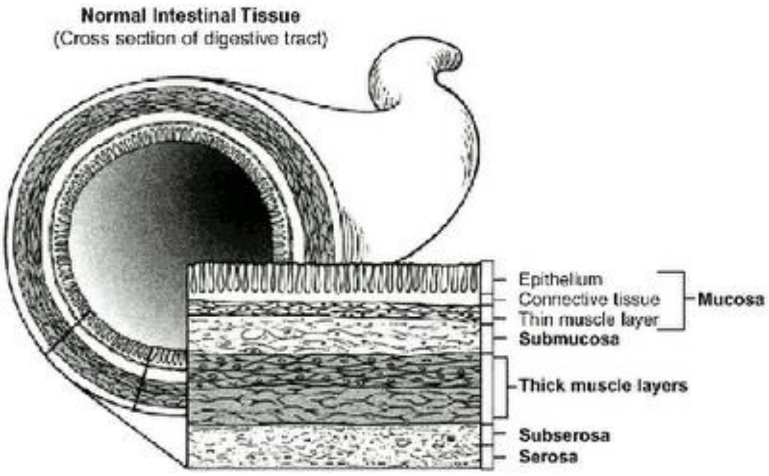


Figure 8.12 A cross-section drawing of the colon demonstrating the mucosa layer in the luminal surface of the colon.

Immunostaining is a key technique widely used to identify the absence or presence of a protein, its tissue distribution, subcellular localisation and changes in the expression, and is heavily dependent on a sensitive and specific antibody. As demonstrated in previous sections, the commercial antibody to human SHROOM2 detects a ~170kDa band on Western Blotting that is reduced with siRNA knockdown, but also detects multiple other non-specific bands of varying molecular weights. Antibodies from other commercial companies were even less effective (data not shown), hence we sought to generate specific antibodies in rabbits with two separate antigens consisting of amino acids sequences from distinct regions of SHROOM2 (Figure 8.20). Only the AQA sera detects SHROOM2 along with several non-specific bands (Figure 8.21A). Dilutions of the AQA sera from rabbit 51 were performed but this did not improve the specificity of the sera (Figure 8.21B).

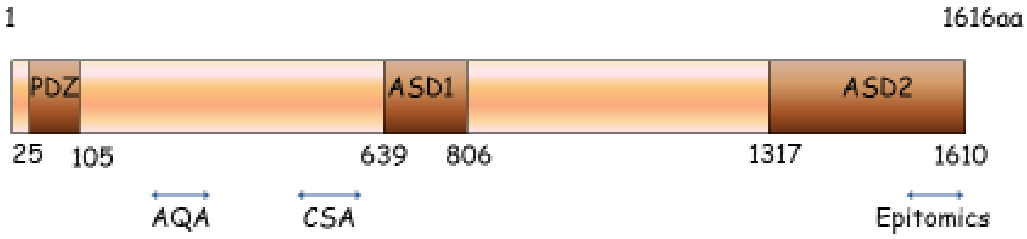


Figure 8.13 A schematic diagram of the human SHROOM2 protein and the protein region targeted by the commercial antibody from Epitomics and custom antibodies AQA and CSA.

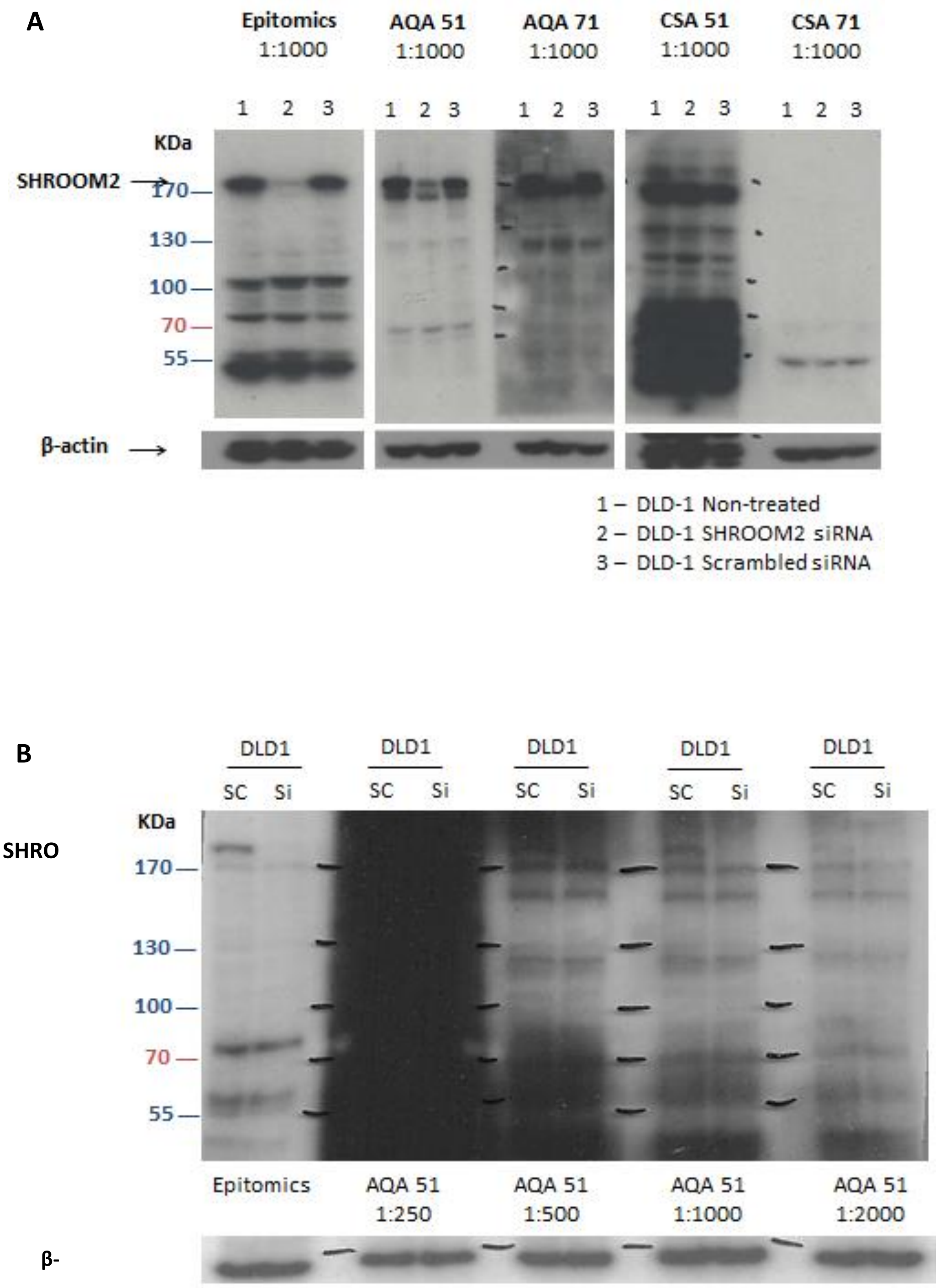


Figure 8.14 A) Western blots of total cell extracts from DLD1 cells with *SHROOM2* knockdown using siRNA1. Commercial (Epitomics) and custom antibodies to SHROOM2 were used for detection of SHROOM2. B) Dilutions of the AQA sera (rabbit 51) with the Epitomics commercial antibody as a positive control.

Without a specific antibody to perform immunostaining, I attempted to characterise the subcellular localisation of *SHROOM2* by subcellular fractionation and Western Blotting. In DLD1 cells, it appears that *SHROOM2* is largely expressed in the cytosolic compartment and the cytoskeletal compartment (Figure 8.22). LS174T is a non-expressing cell line and was used as a negative control. This fits in well with published studies of SHROOM2, where it has been reported to associate with F-actin and is involved in regulating the cytoskeletal organisation and architecture of endothelial cells (Dietz *et al*, 2006). The markers of various subcellular fractions are shown to demonstrate that there is minimal cross-over of the compartments, but this is not without caveats and should be interpreted in combination. For instance, the nuclear marker Lamin-B stains both nuclear and cytoskeletal compartments, as being a nuclear intermediate filament protein it also precipitates into the insoluble cytoskeletal fraction. However, staining with the cytoskeletal marker Vimentin demonstrates that there is unlikely to be contamination of the nuclear fraction with cytoskeletal proteins.

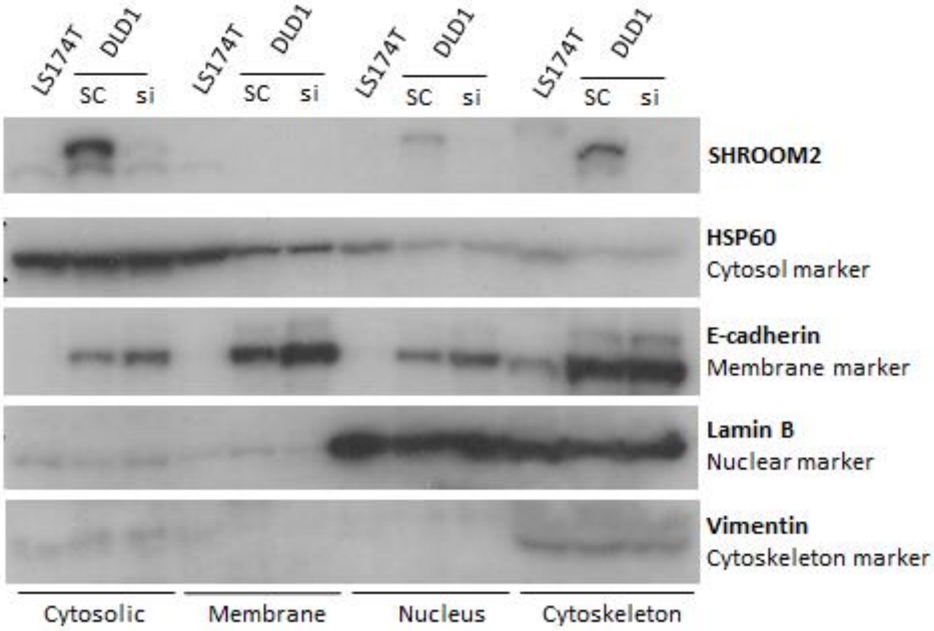


Figure 8.15 Western blots of subcellular fractions extracted from LS174T cells and DLD1 cells with *SHROOM2* knockdown using siRNA1. The commercial antibody (Epitomics) was used for detection of SHROOM2.

To characterise the spatial distribution of SHROOM2 within the colonic normal mucosa, I sought circumferential evidence by examining its expression correlation with various tissue and cell type markers within the expression microarray data of the normal mucosa samples (n=115). Firstly the markers for the three tissue layers of the mucosa were examined in relation to *SHROOM2* expression (Table 8.8), using conventional markers or novel markers recently identified (Pinchuk *et al*, 2010; Powell *et al*, 2011; Roberts *et al*, 2014). Endothelial markers were also included to account for possible inclusion of the submucosa during the tissue harvesting process. Values for each individual probe are shown when there are multiple probes for a given gene. Although the different probes for an individual gene are not always consistent in expression levels and degree of correlation, it is reassuring that in general, they exhibit similar directionality (Figure 8.23). The average gene expression (Table 8.8 and Figure 8.23A) suggests that the tissue sampled consisted of epithelial cells, mesenchymal stromal cells, smooth muscle cells and endothelial cells in decreasing order, which broadly reflects the expected composition of the normal mucosa with minimal submucosal contamination. Several marker genes for all the different cell types were significantly correlated with the expression of *SHROOM2*, but only *CDH1* (E-cadherin), an epithelial cell marker, was positively correlated with *SHROOM2* (Table 8.8 and Figure 8.23B and C), suggesting that *SHROOM2* is more likely to be expressed in epithelial cells than other tissue-types. Interestingly, *SHROOM2* is significantly negatively correlated with the epithelial cell marker *CD44*, which would argue against *SHROOM2* being expressed in the epithelium. However, this negative correlation may reflect the differential expression of *SHROOM2* along the colonic crypt vertical axis as *CD44* is known to be expressed in the crypt base of the murine (Rothenberg *et al*, 2012) and human colon (Dalerba *et al*, 2011).

Tissue/cell type	Gene symbol	Log_2 average expression	Nominal *p*-val	Spearman rho
Epithelial	*CDH1*	10.97	0.0006***	0.311
	EPCAM	12.63	0.0661	0.171
	CD44	9.98	0.0145*	-0.226
	CD44	7.40	0.0887	-0.158
	CD44	7.10	0.1183	-0.145
Mesenchymal stromal	*ACTA2*	11.69	0.9300	-0.008
	THY1	7.58	0.2100	-0.116
	DES	7.07	0.7500	0.030
	VIM	10.22	0.0400*	-0.187
	VIM	10.98	0.1400	-0.136
Smooth muscle	*CALD1*	8.08	0.2067	-0.118
	CALD1	7.82	0.5392	-0.057
	CALD1	9.21	0.9277	-0.008
	SMTN	6.59	0.0340*	-0.196
	SMTN	7.64	0.1810	-0.124
	SMTN	6.71	0.0354*	-0.195
Endothelial	*PECAM1*	7.91	0.0860	-0.159
	CD34	7.43	0.0031**	-0.272
	CD34	6.58	0.6000	0.049
	CD34	8.08	0.7300	0.032
	CD34	6.51	0.7500	0.030
	MCAM	6.54	0.9800	0.002
	ENG	7.50	0.0480*	-0.184

Table 8.16 Cell-type specific markers and their correlation with *SHROOM2* in the normal colorectal mucosa. Values are shown for each probe when there are multiple probes for a gene.

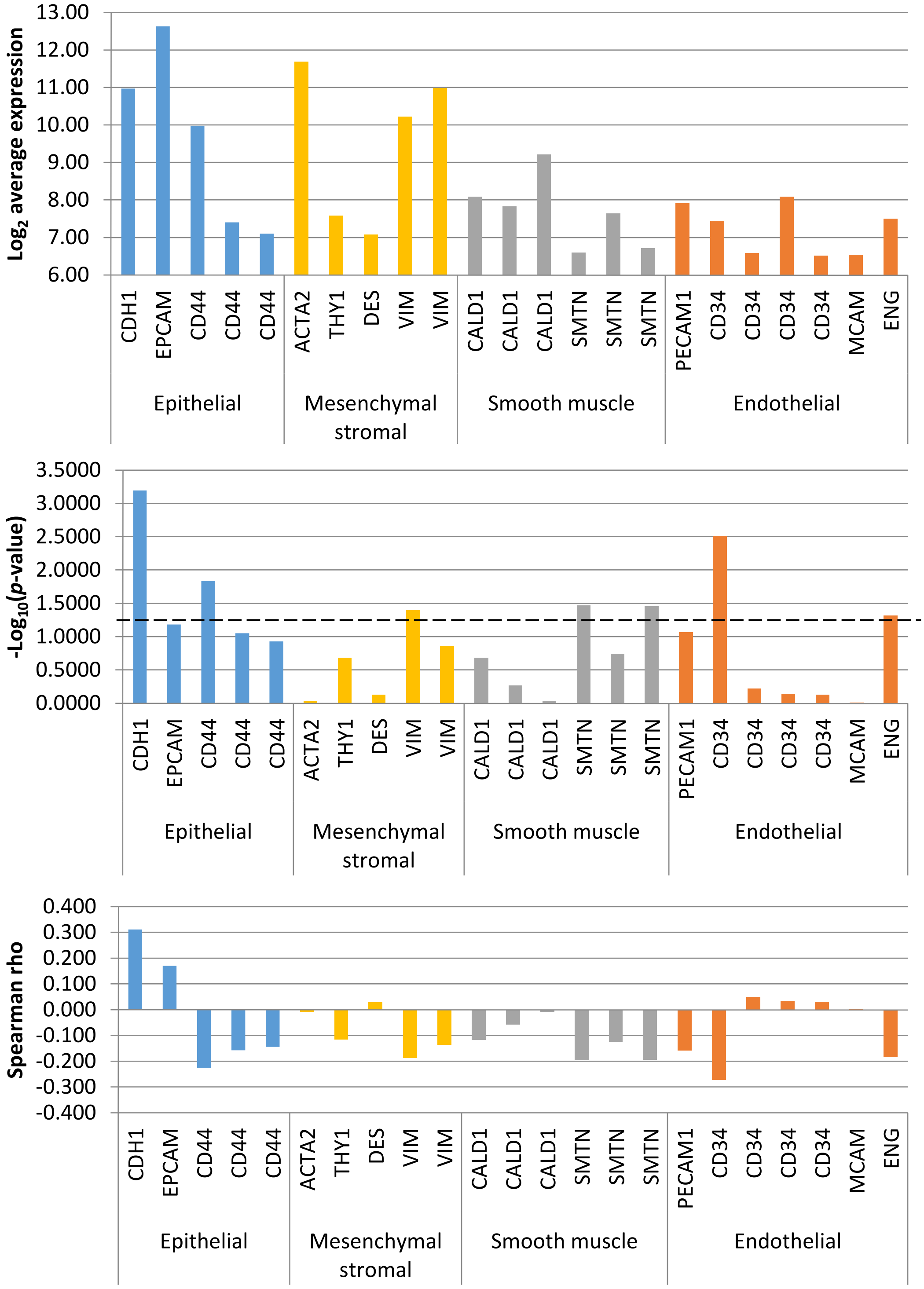


Figure 8.16 A) The average expression of tissue/cell type specific markers. B) the -log$_{10}$(*p*-value) of the correlation between *SHROOM2* expression and marker genes. Dotted line represents a *p*-value of 0.05. C) Spearman rho of the correlation between *SHROOM2* and marker genes. Values are shown for each probe when there are multiple probes for each gene.

There is increasing evidence of heterogeneity of cells within the human colonic epithelium, with different protein markers and transcriptional signatures reflecting their lineage, differentiation stage and functional status (Dalerba *et al*, 2011). By inference, *SHROOM2*'s association with these markers may provide insight into its role within the colonic epithelial crypt. Useful markers that are frequently expressed in a mutually exclusive way include genes encoding lineage-specific markers such as *CA1* for enterocytes, *MUC2* for goblet cells, and *LGR5* for the immature compartment. From a differentiation point of view, there are also transcriptional programs that characterises "top-of-the-crypt" mature, differentiated enterocytes and "bottom-of-the-crypt" cell populations which include compartments characterised by genes linked to goblet cells and genes that are expressed in the progenitor cell compartment of the mouse small intestine. (Dalerba *et al*, 2011; Merlos-Suarez *et al*, 2011). Functionally, the expression of proliferation markers can also be examined; these are generally restricted to the bottom of the crypts.

SHROOM2's appears to be positively correlated with CA1 and not MUC2 or LGR5 (Table 8.9 and Figure 8.24), suggesting that it is expressed mainly in enterocytes. Overall, there are more significantly positive correlations between *SHROOM2* and genes that are highly expressed by "top-of-the-crypt" differentiated enterocytes, as compared to genes that are expressed by "bottom-of-the-crypt" cells. Although there is significant correlation with some of the proliferative markers, these correlations are negative and hence consistent with expression in non-proliferative or mature differentiated enterocytes. There are also positive correlations with two of the genes expressed in the progenitor cell compartment (*RGMB* and *PTPRO*), which may relate to *SHROOM2*'s subcellular localisation and function rather than crypt distribution, as there was no correlation with the other markers of stem-ness (*LGR5* and *ASCL2*). This is rather intriguing as they are both plasma membrane associated proteins that have a regulatory role in cellular growth, differentiation and cell cycle progression - RGMB is a glycosylphosphatidylinositol (GPI)-anchored protein that potentiates BMP signalling (Halbrooks *et al*, 2007), whereas PTPRO is a protein tyrosine phosphatase localised to the apical surface of polarised cells that interferes with cell cycle progression (Motiwala *et al*, 2004). Alternatively, these could be false positive results given the number of genes tested for a correlation with *SHROOM2*.

Distribution	Lineage/Function/ Differentiation	Gene symbol	Log_2 Average expression	Nominal *p*-val	Spearman rho
Crypt-top	Mature enterocytes	*SLC26A3*	13.19	0.0001***	0.351
		CA1	12.32	0.0248*	0.208
		CA2	12.96	0.0331*	0.197
		CA2	12.35	0.1256	0.142
		MS4A12	10.82	0.0047**	0.260
		AQP8	11.20	0.0008***	0.307
		CD177	7.12	0.0067**	0.250
		CD177	7.53	0.0069**	0.249
		KRT20	9.05	0.0171*	0.220
		KRT20	11.94	0.0501	0.182
Crypt-bottom	Goblet	*MUC2*	11.52	0.7838	0.026
		TFF3	13.23	0.2982	-0.097
		SPDEF	8.04	0.0703	-0.168
		SPINK4	10.38	0.1295	-0.141
	Stem/Progenitor	*LGR5*	6.56	0.9477	0.006
		OLFM4	11.16	0.0605	-0.174
		OLFM4	7.52	0.0552	-0.178
		ASCL2	7.80	0.6126	-0.047
		RGMB	6.64	0.1243	-0.143
		RGMB	7.87	0.0242*	0.209
		RGMB	6.58	0.0181*	0.219
		PTPRO	7.41	0.0365*	0.194
		PTPRO	7.39	0.0330*	0.197
	Proliferative	*MKI67*	6.76	0.0115*	-0.233
		TOP2A	8.65	0.0227*	-0.211
		BIRC5	6.49	0.1166	-0.146
		BIRC5	7.20	0.5412	-0.057

Table 8.17 Genes characteristically expressed by subpopulations of epithelial cells and their correlation with *SHROOM2* in the normal colorectal mucosa. Values are shown for each probe when there are multiple probes for a gene.

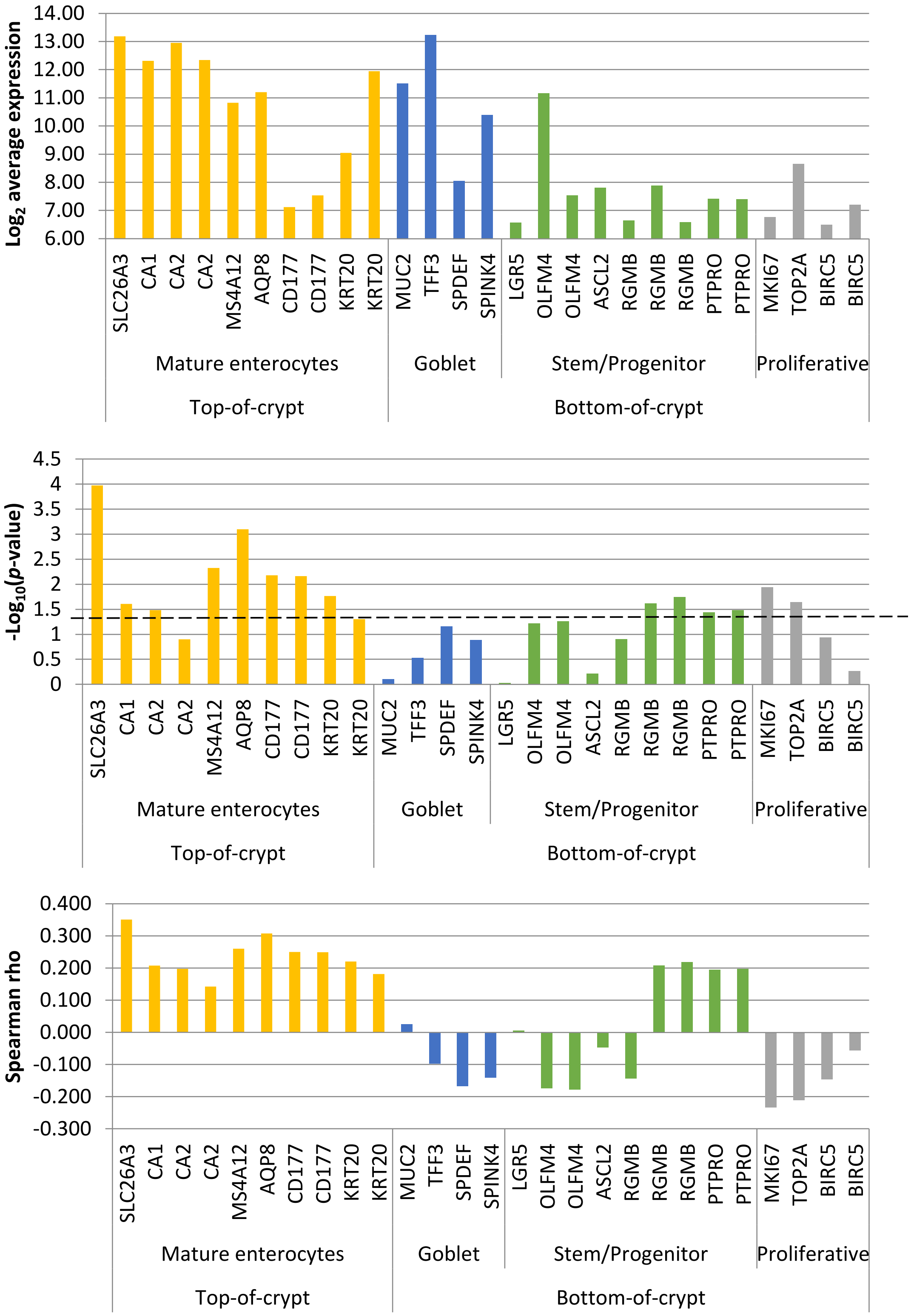


Figure 8.17 A) The average expression of genes characteristic of colonic epithelial subpopulations. B) the -log$_{10}$(*p*-value) of the correlation between *SHROOM2* expression and marker genes. Dotted line represents a *p*-value of 0.05. C) Spearman rho of the correlation between *SHROOM2* and marker genes. Values are shown for each probe when there are multiple probes for a gene.

In summary, by Western Blotting of subcellular fractions, SHROOM2 appears to be a cytosolic protein that associates with the cytoskeleton in DLD1 colorectal cancer cell line. By examining the correlation of *SHROOM2* with gene-expression markers of the different cell populations within the colonic mucosa, it was deduced that *SHROOM2* is mainly expressed in mature enterocytes of the epithelial layer at the top of the colonic crypts. An effective antibody to SHROOM2 would provide the validation required to substantiate these inferential findings.

8.4 Discussion

The data presented in Chapter 6 and 7 has shown that the Xp22.2 CRC risk locus demonstrates eQTL activity targeting the cis-gene *SHROOM2*, with evidence to suggest that the causal indel variant at the distal promoter region influences *SHROOM2* transcription by modulating NF-Y transcription factor binding. By association, *SHROOM2* is implicated as a candidate susceptibility gene and this chapter presents preliminary data on the function of *SHROOM2* in human colonic epithelial cells.

Prior to any functional assays of *SHROOM2*, it is important that the presence of transcript isoforms are first identified in the tissue type of interest, as it has been reported that non-coding genetic variation may influence disease risk by altering levels of expression and splicing architecture of mRNA transcripts (Graham *et al*, 2006; Zhang *et al*, 2009). Furthermore, characterisation of the relevant transcript isoforms will ensure that any observed phenotype is accurately attributed to a specific transcript isoform. This is particularly important for *SHROOM2* as there are four reported transcripts and the HT12 microarray probe used for the eQTL analysis only detects the large canonical transcript *T-001*. Hence, it is reassuring that it is the only transcript that is expressed in the primary normal mucosa and tumour tissue and most of the cell lines used for functional assays (8.3.2). There is some evidence of tissue-specific alternative splicing as the other protein coding transcript isoforms are present in cell lines derived from other tissue types such as the retinal pigment epithelium, breast cancer and prostate tissue, suggesting different or additional functional roles for *SHROOM2* in extra-colonic tissue.

The data presented in the next part of the chapter focuses on the transient knockdown of *SHROOM2* in tumour cell lines with RNA interference, and the effect on tumourigenicity assays such as growth curves and scratch-wound assays. Traditional cancer cell lines have the advantage of being accessible, easily manipulated and replicable, hence ideal for in-vitro cell biology experiments where near-complete control of the environment and the existence of a single cell type are desirable. However, not only do these cell lines harbour genetic aberrations, they have been subjected to gross manipulation during the process of creating cell lines and are kept under artificial conditions, therefore are not necessarily an accurate representation of the natural cellular state in vivo. The preliminary work presented here did not show a detectable change in the growth curves and scratch-wound assays with transient RNAi of *SHROOM2*, however, these negative results may reflect the major caveat that accompanies tumour cell lines, where overbearing mutator phenotypes in proliferative signalling and cell-cycle check points would displace or override any subtle effect of a lower-risk gene. Alternatively, it may represent a requirement for further experimental optimisation. For example, cell counting is a rough-and-ready technique and more specific readouts such as viability assays, proliferative and apoptotic markers, cell cycle phase markers may be more informative. With regards to the scratch wound assay, the cell line controls used only closed the scratch-wound gap minimally, with a closure of only ~25% after 24 hours. This suggests there may be factors impeding the migration of the cells in general, and any effect of gene depletion may have been masked. Further optimisation of this system will be required to robustly demonstrate the role of *SHROOM2* in cell migration, for instance, knockdown of a positive control such as *FAK*, titration of serum in the growth medium and/or cell confluence, a longer time-course, or use of chemoattractants and gradient chambers. An experimental design utilising repeated measures analyses could also be used in future experiments to allow longitudinal monitoring and analysis. This will improve the power of detecting changes and order effects, and may reveal more subtle changes in cellular phenotypes associated with *SHROOM2* depletion.

A further consideration for future functional work is stable knockdown/knockout of *SHROOM2* or overexpression vectors, as certain cellular phenotypes may not be easily observed and quantified within the short time frame of transient siRNA depletion. Alternative models such as mouse models, ex-vivo three-dimensional organoid culture or normal colonic epithelial cell lines would offer a relatively normal physiological platform investigate gene functions, albeit more challenging and costly/labour-intensive. Stable genome editing using CRISPR-Cas9 technology has recently been successfully performed for driver pathway mutations in human colonic organoids as a model of colorectal cancer (Matano *et al*, 2015), and may ultimately allow more definitive phenotype characterisation for candidate susceptibility genes such as *SHROOM2* before progressing to dissect the subtle phenotypes associated with gene dosage.

Co-expression analysis implicates that *SHROOM2* is expressed in the mature differentiated enterocytes at the top of the colonic epithelial crypts, and gene ontology analysis of cell line RNAi and correlated genes in the normal mucosa is suggestive of a role in cell cycle regulation. It is known that the transcription factor implicated in its transcriptional control, NF-Y, also regulates transcription of various genes related to the cell cycle (Elkon *et al*, 2003; Caretti *et al*, 2003) and that co-regulated genes often share biological functions. This indirectly lends further support to the suggestion that *SHROOM2* may exert its tumour suppressive effects by influencing the cell cycle progression, which can in turn, influence proliferation, differentiation and apoptosis. It would therefore be of considerable interest to design and perform experiments that would directly implicate *SHROOM2* in the regulation of cell cycle, such as flow cytometry cell cycle analysis or FUCCI (Fluorescence Ubiquitin Cell Cycle Indicator) cell cycle reporter vectors.

The expression of many genes as well as eQTLs effects appear to be tissue- or cell-type specific. As demonstrated in 6.3.2, not only is *SHROOM2* lowly expressed in PBMCs compared to normal mucosa, the rs5934683-*SHROOM2* eQTL is also not detectable. Given the heterogeneity of cell-types within the normal mucosa as well as the expression gradient along the epithelial crypt axis, it is conceivable that *SHROOM2* expression and/or the eQTL effect are selectively present in a similar fashion. Risk alleles and their target genes may act in a non-cell autonomous fashion and therefore may exert their effect through other cell types that act upon the target cells, it is therefore crucial to further understand the spatial distribution of *SHROOM2* within the colonic mucosa. This is particularly so as it is well-recognised that the mesenchymal stroma plays a key paracrine signalling role in maintaining the epithelial crypt architecture, and that cancers are thought to arise from the dysregulation of the crypt-base stem cell niche which harbours the cell of origin of colorectal cancer - the ‘bottom-up’ hypothesis. The co-expression analysis of *SHROOM2* and cell-specific marker genes in the normal mucosa indicates that

SHROOM2 expression is correlated with markers of crypt-top mature enterocytes, suggesting that *SHROOM2* is mainly expressed in this cell sub-type. This may or may not be representative of the eQTL effect. As there is rapid cell turnover and the majority of differentiated cells are shed into the lumen within 5days, perturbations in this crypt compartment are in theory less likely to initiate neoplasia. Interestingly, a recent mouse model study of hereditary mixed polyposis syndrome (HMPS) demonstrated evidence for the alternative 'top-down' hypothesis of tumour formation, where aberrant epithelial expression of *GREM1* promoted the persistence and/or reacquisition of stem cell properties in *LGR5*-negative cells that have exited the stem cell niche, allowing cells outside the crypt-base stem cell niche to form ectopic crypts and act as tumour progenitors (Davis *et al*, 2014). Sporadic traditional serrated adenomas, which are characterised by ectopic crypt foci, were also shown by the authors to express epithelial *GREM1*, suggesting a similar underlying mechanism. In another study, activated NF-κB induced mucosal inflammation in combination with constitutive epithelial Wnt signalling was shown to promote the initiation of neoplasia from cells situated outside the crypt base stem cell niche (Schwitalla *et al*, 2013). Given that these studies demonstrate that the top-down model of tumourigenesis may indeed fit some subtypes of inherited and sporadic colorectal cancers, it would be of great interest to refine the tissue localisation of *SHROOM2* and the eQTL effect. Apart from immunostaining techniques, RNA-FISH could be used to demonstrate the spatial distribution of *SHROOM2* transcripts. Alternatively, disaggregation of the intestinal epithelial crypts from the underlying stromal tissue using enzymatic or non-enzymatic methods can be performed on freshly sampled colonic mucosa prior to extraction of RNA/protein. During the process of normal mucosa sampling, tumour tissue has also been collected and fresh frozen, and may shed light on the role of *SHROOM2* in tumourigeneis. *SHROOM2* expression and the eQTL effect could be examined, as well as any associations with driver pathway mutations, chromosomal abberations, microsatellite instability and epigenetic changes (e.g. CIMP phenotype).

In the context of human case-control studies, the study that would cement *SHROOM2*'s role in CRC risk is to quantify *SHROOM2* expression in the normal mucosa of cases versus controls, ideally in a prospective manner. This is challenging to say the least, but is possible in the longer term with the increasing realisation that normal tissue repositories are a vital resource in understanding disease mechanisms. In the larger scheme, this would also be an invaluable resource in facilitating the amalgamation of genetics, transcriptomics and proteomics of normal tissue states that would complement the ongoing work in disease states, and to discover and study biomarkers of disease predisposition and clinical outcomes.

In summary, this chapter presents preliminary data on the functional role of *SHROOM2*, the target gene of a GWAS-associated common variant that is implicated as a colorectal cancer susceptibility gene. Subcellular localisation studies and co-expression analysis of the normal mucosa suggests that it is a cytoplasmic protein that associates with the cytoskeleton, and is likely to be mainly expressed in the crypt-top mature enterocytes. RNA interference studies suggests a role in the regulation of cell cycle progression, and further understanding of role of *SHROOM2* would prove invaluable in understanding the contributory pathways to CRC carcinogenesis, and ultimately inform the rational development of preventative/therapeutic strategies for colorectal cancer.

Chapter 9

Functional characterisation of the gene-environment (plasma 25-hydroxyvitamin D) interaction at the 16q22.1 risk locus

9.1 Introduction

The eQTL analysis of colorectal mucosa and PBL has provided evidence of regulatory function for approximately half of the CRC risk SNPs (Chapter 6). Further functional studies of the Xp22.2 locus validates this eQTL association (Chapter 7), providing proof of principle on a molecular level. However, there is still a large proportion of risk variants whose functions and target genes are unaccounted for.

There is emerging evidence of gene-environment interactions on cancer risk in the context of low-penetrance genetic susceptibility polymorphisms, for instance, parity and alcohol consumption influencing breast cancer genetic risk conferred by common alleles (Nickels *et al*, 2013), and common genetic variation modifying the protective effect of NSAID/aspirin use in colorectal cancer (Nan *et al*, 2015). Although the underlying molecular mechanisms have yet to be identified, these studies are suggestive that the *function* of common variants may also be modified by non-genetic factors. Hence, it is conceivable that the eQTL effects of CRC risk variants may not be fully appreciable under steady-state conditions, in other words, their association with the expression of target genes may be modifiable by perturbations of cellular pathways that are influenced by lifestyle/environmental factors.

A recent study by the Colon Cancer Genetics Group (Zgaga *et al*, unpublished) investigated whether vitamin D levels modified the risk conferred by the 25 common variants associated with CRC. This investigation stemmed from the implication of vitamin D deficiency as a possible risk factor in the aetiology of colorectal cancer, where higher vitamin D intake, higher serum 25-hydroxyvitamin D (25-OHD) and residence in regions with strong UVB radiation were associated with lower CRC risk (Gorham *et al*, 2005; Giovannucci 2009; Gandini *et al*, 2011) and cancer mortality

(Robsahm *et al*, 2004, Tretli *et al*, 2012). Interestingly, this study demonstrated a statistically significant 2-way interaction between plasma 25-OHD and rs9929218 at the 16q22.1 locus (p=0.004) (Figure 9.1). In other words, the effect of rs9929218 on CRC risk was modified by the levels of plasma 25-OHD (Table 9.1). This is the first study to implicate an interaction between a known CRC risk variant and an environmental factor; if true, this could have a significant impact on public health strategies in CRC risk stratification, screening and prevention. Hence, there is much value in pursuing an understanding of the functional mechanism that mediates this gene-environment interaction observed in population studies.

The rs9929218 SNP (chr16:68820946) resides within intron 2 of *CDH1* that codes for E-cadherin, a protein that plays a crucial role in epithelial cell-cell adhesion and tissue architecture maintenance. It has been previously implicated in colorectal cancer (as reviewed by Tsanou *et al*, 2008), and its reduced expression is known to be associated with invasive potential and poor prognosis in various cancers. Although intron 2 of *CDH1* is known to contain cis-regulatory elements for its transcription (Stemmler *et al*; 2005), the expression analysis of CRC risk variants in normal mucosa and PBMC (Chapter 6) did not reveal any evidence to suggest a relationship between rs9929218 genotype and *CDH1* expression. Similarly, ChIP-seq studies of the vitamin D receptor (VDR) does not show vitamin D response elements at this locus (Ramagopalan *et al*, 2010;) and independent bioinformatics analysis does not support VDR binding at this locus (Zgaga *et al*, unpublished). Instead, the in-silico analysis strongly suggests a putative FoxO binding site at rs9929218, with a 10-fold increase in FoxO binding affinity associated with the A allele. This is intriguingly as there is evidence in the literature to indicate that VDR associates directly with FoxO proteins and their regulators, and that vitamin D treatment induces post-translational modification of FoxO proteins, enhancing their binding to the promoters of target genes (An *et al*, 2010). There is also evidence suggesting that FoxO3a is involved in the regulation of E-cadherin expression in urothelial cancer cells (Shiota *et al*, 2010).

From this, it can be postulated that the observed interaction between 25-OHD and rs9929218 on CRC risk is mediated by *VDR*'s ligand-dependent non-genomic actions, whereby it modulates the activity of FoxO proteins on cis-regulatory elements of *CDH1*. This chapter aims to investigate this hypothesis by utilising the expression data derived from the human colonic mucosa and PBMC, as well as measurements of matched serum 25-OHD. To investigate further whether the rs9929218 genotype influences the induction of *CDH1* expression, the in-vivo expression analysis will be complemented by the in-vitro vitamin D treatment of CRC cell lines and human colonic organoids.

One of the biggest limitations of studies utilising single 25-OHD measurements in observational studies of cancer incidence and mortality is that 25-OHD is frequently measured after the diagnosis. Determining the direction of causality is challenging as 25-OHD levels may have decreased as a result of illness or treatment. Indeed, decreased circulating 25OHD concentration has been reported after elective knee surgery (Reid *et al*, 2011) and cardiopulmonary bypass (Krishnan *et al*, 2010). As the majority of vitamin D measurements used in this chapter were taken post-operatively at varying intervals (ranging from 1to 468 days), they may not accurately reflect the vitamin D status at the point of mucosa sampling given the reported changes that accompany major surgery. Therefore, it is of importance to characterise these changes after abdominal surgery for large bowel resections. Understanding how circulating 25-OHD fluctuates peri- and post-operatively will not only contribute towards robust statistical analysis of its association with disease and expression phenotypes, it will also offer insight into the regulation and homeostasis of vitamin D and inform the design of future studies investigating its role in the development of complex disease traits.

25-OHD TERTILE	N	GA vs. AA			GG vs. AA		
		OR	95% CI	*p*	OR	95% CI	*P*
1	1357	0.92	0.61-1.41	0.71	0.97	0.64-1.46	0.87
2	1449	1.27	0.85-1.9	0.24	1.49	1.01-2.21	0.044
3	1410	1.83	1.15-2.91	0.01	2.35	1.49-3.7	0.0002

Table 9.1 Association between rs9929218 and colorectal cancer for different tertiles of May-standardised 25-OHD. The cut-offs for the 25-OHD tertiles (T1, T2 and T3) were: 0-8.3, 8.4-14.5 and >14.6 ng/ml (Zgaga *et al*, unpublished).

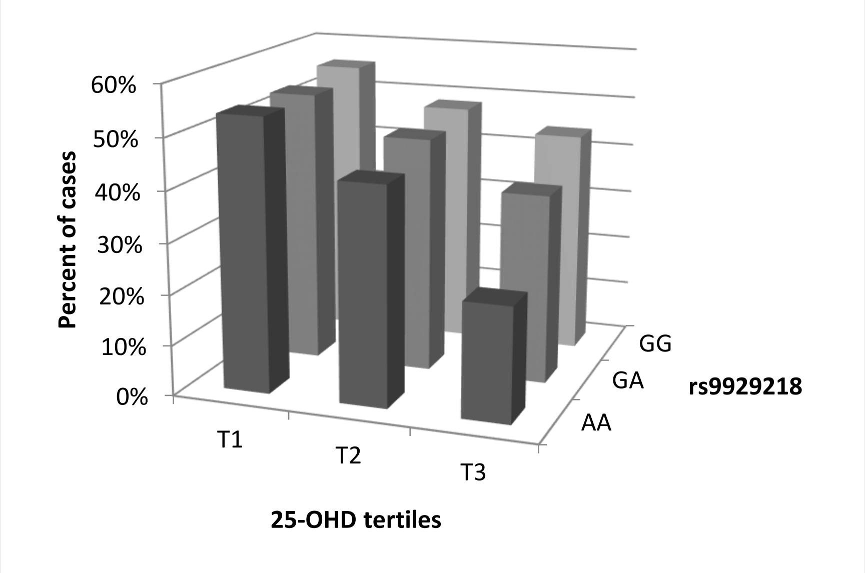


Figure 9.1 The proportion of colorectal cases in subgroups based on rs9929218 genotype and 25-OHD tertiles (Zgaga *et al*, unpublished).

9.2 Methodological overview

9.2.1 Study subjects and biological material

To investigate the factors influencing *CDH1* expression in-vivo, normal colonic mucosa (n=115) and matched PBMC (n=59) were collected from patients undergoing large bowel surgery as described in 6.2.1. In a subset of these patients (n=83), blood was also collected post-operatively for the quantification of circulating 25-OHD. Plasma was isolated from peripheral blood as described in 2.1.2.

A different cohort of patients undergoing large bowel surgery for CRC were recruited for the serial sampling study of serum 25-OHD (n=40) (Table 9.2). Six serum samples were obtained from these patients at the following time points – pre-operatively (3-19 days before surgery), 1-2 days post-op, 3-5 days post-op, 6-8 days post-op, first outpatient follow-up (30-120 days post-op), and second outpatient follow-up (>162 days post-op).

MD	AGE	GENDER	AJCC STAGE	OPERATION
11015	74	M	2	Laparoscopic
11028	82	M	1	Open
12692	63	M	1	Laparoscopic
12755	76	F	2	Open
12777	73	F	3	Open
12781	81	F	3	Laparoscopic
12785	57	F	1	Open
12788	73	F	1	Laparoscopic
12789	88	M	3	Open
12794	86	F	3	Laparoscopic
12796	76	F	2	Laparoscopic
12797	75	M	2	Laparoscopic
12798	61	M	1	Laparoscopic
12800	65	F	2	Laparoscopic
12802	71	F	1	Laparoscopic
12804	46	M	1	Laparoscopic
12805	49	F	2	Laparoscopic
12807	68	M	2	Open
12808	69	M	1	Open
12811	78	F	3	Laparoscopic
12812	52	M	1	Open
12813	78	M	1	Open
12815	65	M	2	Laparoscopic
12822	73	F	3	Open
12824	65	F	1	Laparoscopic
12826	83	F	2	Open
12857	71	F	1	Laparoscopic
12873	64	M	2	Laparoscopic
12874	65	M	2	Laparoscopic
12876	76	M	2	Laparoscopic
12882	81	M	2	Open
12887	65	M	1	Open
12890	85	M	2	Open
12893	81	M	2	Laparoscopic
12897	52	M	1	Laparoscopic
12898	85	M	3	Laparoscopic
12903	78	M	2	Open
12904	60	F	1	Laparoscopic
12906	49	M	1	Open
12919	76	M	2	Laparoscopic

Table 9.2 Characteristics of the 40 study subjects – age, gender, AJCC stage of CRC and the type of large bowel surgery (open or laparoscopic).

9.2.2 Gene expression levels and variant genotypes

The expression of the genes of interest was extracted from the microarray data after normalisation, batch correction and log transformation as described in 2.3.6. Genotypes of the SNPs of interest were obtained from the HumanOmni5M-4v1_B BeadChip Arrays (Illumina, USA) as described in 6.2.3.

9.2.3 Calcitriol treatment of cell lines

Cell lines were cultured as described in 2.2 until 50% confluence prior to calcitriol treatment. Calcitriol (Sigma-Aldrich) was reconstituted in 100% ethanol and a final concentration of 100nM were used. The equivalent volume of 100% ethanol was used as the negative control, which equates to 1% v/v ethanol. 10% charcoal-stripped fetal bovine serum (Life Technologies) was used during calcitriol treatment to eliminate lipophilic material that contain vitamin D metabolites that may mask or falsely elevate the effect of calcitriol treatment.

9.2.4 Calcitriol treatment of human organoids

Human colonic organoid culture was carried out as described in Sato *et al*, 2011 using epithelial crypts dissociated from the colonic tissue harvested from fresh surgical specimens. In brief, epithelial crypts are dissociated from the stroma using 25mM EDTA and mechanical pipetting. After washing, the crypts were resuspended in Matrigel (BD Bioscience) at 200 crypts per 50µL of Matrigel in each well (24-well plate). 500µL of culture medium was placed in each well after the Matrigel has solidified, and culture medium was replaced every 2 days. The organoids were incubated at 37ºC in a humidified incubator (95% O2, 5% CO2). At day 5 in culture when crypt budding started to occur, the organoids were treated with 100nm calcitriol (Sigma-Aldrich) or the 1% v/v ethanol negative control for 16 hours.

The following constitutes the organoid culture medium:

REAGENT	SOURCE	FINAL CONCENTRATION
Advanced DMEM/F12	Invitrogen™, Life Technologies	1x
Glutamax	Invitrogen™, Life Technologies	2mM
Hepes	Invitrogen™, Life Technologies	10mM
Bsa	Sigma-Aldrich	0.1%
Penicillin/Streptomycin	In-house technical services	100U/130μg per ml
N-acetylcysteine	Sigma-Aldrich	1mM
N2	Invitrogen™, Life Technologies	1x
B27	Invitrogen™, Life Technologies	1x
Gastrin I	Sigma-Aldrich	10nM
Nicotinamide	Sigma-Aldrich	10mM
A83-01	Tocris	500nM
SB202190	Sigma-Aldrich	10μM
Noggin (mouse recombinant)	Peprotech	100 ng/ml
Epidermal Growth Factor (mouse recombinant)	Invitrogen™, Life Technologies	50ng/ml
R-Spondin (mouse recombinant)	R&D	1 μg/ml
Wnt-3a (mouse recombinant)	R&D	100ng/ml

Table 9.3 Details of the reagents used for the human organoid culture medium.

9.2.5 qRT-PCR

RNA from cell lines and human organoids was extracted using the Ambion® RiboPure™ RNA extraction kit (Life Technologies) as per the manufacturer's protocol. DNAse treatment, cDNA synthesis and qRT-PCR of the genes of interest were performed as described in 2.3, using the Taqman® Gene Expression assays listed in Table 9.4. Genes of interest were normalised to the reference gene *ACTB.*

Gene	Assay ID	Probe sequence
CDH1	Hs01023895_m1	AAGGTGCTCTTCCAGGAACCTCTGT
VDR	Hs01045844_m1	TGAAGGAGTTCATTCTGACAGATGA
CYP24A1	Hs00167999_m1	GCGGTGGAAACGACAGCAAACAGTC
CYP3A4	Hs00604506_m1	ATTTTGTCCTACCATAAGGGCTTTT
ACTB	Hs99999903_m1	CCTTTGCCGATCCGCCGCCCGTCCA

Table 9.4 Taqman gene expression assays used in the quantification of gene expression in the calcitriol-treated cell lines and human organoids.

9.2.6 Measurement of circulating 25-OHD

Circulating 25-OHD was quantified as a measure of vitamin D status. Total plasma or serum 25-OHD (25-OHD$_2$ and 25-OHD$_3$) was measured by the liquid chromatography-tandem mass spectrometry (LC-MS/MS) method by the Clinical Biochemistry department, Glasgow Royal Infirmary, following standard protocols and quality control procedures (Knox *et al*. 2009). More details about this method can be found elsewhere (Knox *et al*. 2009; Wallace *et al*. 2010). Levels <8nmol/L were undetectable and randomisation was performed based on the distribution of the other samples in the cohort (Figure 9.2). May-adjusted 25-OHD concentrations were used as described in Zgaga *et al*, 2011. To minimise the confounding effects of the season and subsequently daylight length, 25-OHD levels were standardised to the month of May, to remove the effect of the month when blood was sampled on 25-OHD levels (adjusted values of <0nmol/L were re-coded as 0nmol/L). As the majority of samples were considered clinically deficient in 25-OHD (<30nmol/L), the levels were categorised in tertiles for the gene-environment analysis. The cut-offs for the 25-OHD tertiles were: 0-12.5nmol/L, 12.6-23nmol/L and >23nmol/L.

9.2.7 Statistical analysis

All models were adjusted for age, gender and site of sampling. To test for two-way interactions, linear regression analysis was performed, modelling both the main effects and the interaction for the selected SNPs, genes or serum 25-OHD. Where more than one probe was present for a gene, the expression of each probe was analysed individually. Correlation between individual gene expression in the normal mucosa was performed using Spearman correlation as a non-parametric measure of statistical dependence between two variables.

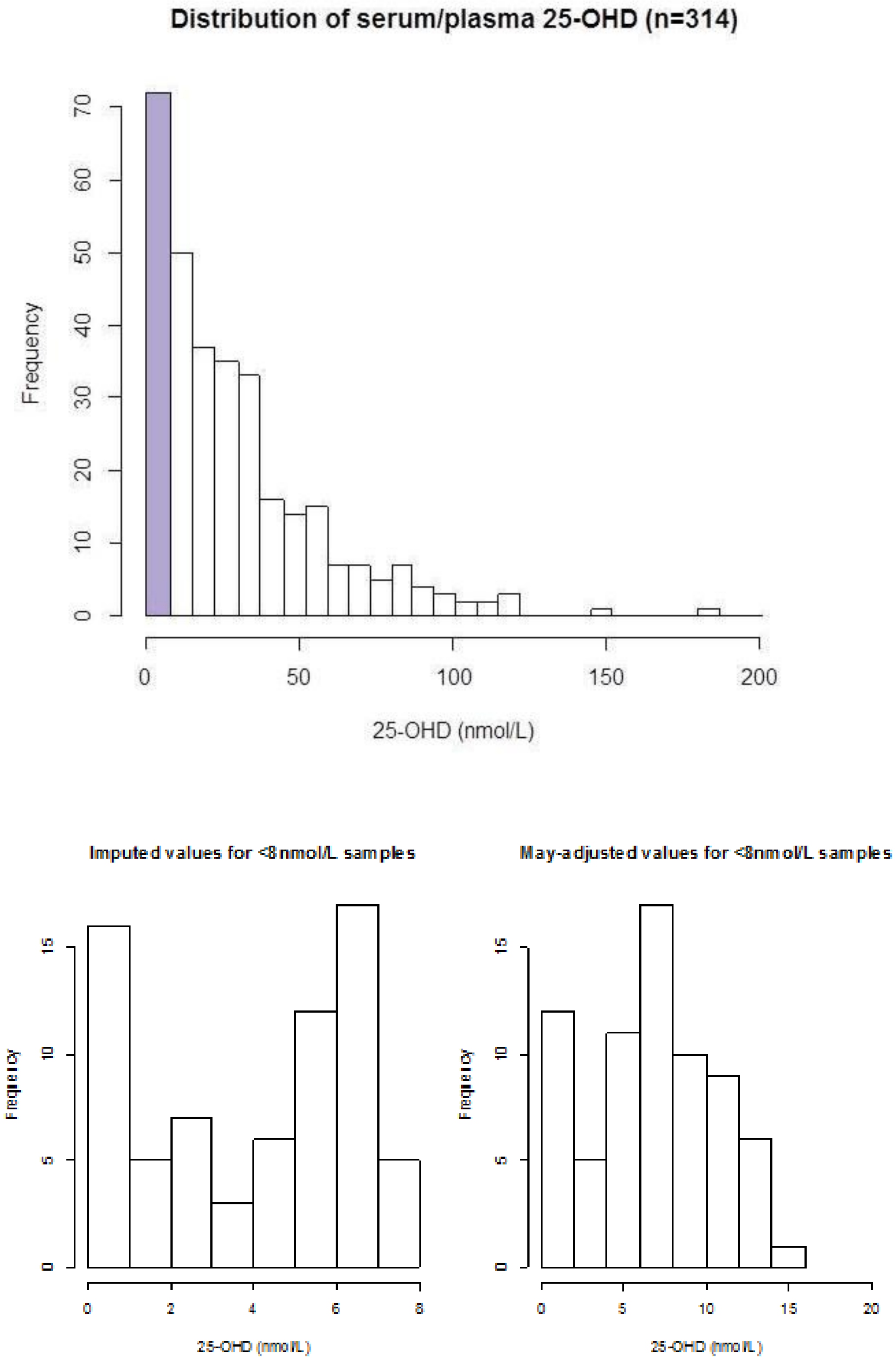


Figure 9.2 Distribution of serum/plasma 25-OHD used in this study (top panel). Samples <8nmol/L (highlighted in purple, n=71) were undetectable by LC-MS/MS and imputation was performed. The distributions of imputed and subsequent May-adjusted values for these samples are shown in the bottom panel.

9.3 Results

9.3.1 Expression of *CDH1* is independently associated with *VDR, CYP3A4* and *FOXO* transcription factors

Vitamin D exerts its biological effects primarily by activating the vitamin D receptor (*VDR*). Upon ligand activation, this nuclear hormone receptor forms a heterodimer with the retinoid-X receptor and binds to vitamin D response elements (VDRE) on DNA, facilitating the recruitment of protein complexes that are essential for transcriptional modulation. There is also evidence of transcriptional autoregulation of *VDR* by the active vitamin D metabolite calcitriol (1,25-dihyroxyvitamin D3) using ChIP anlaysis (Zella *et al*, 2006). Hence, its expression is useful as a marker of vitamin D transcriptional activity and by extension, a proxy of cellular vitamin D status. *CYP3A4* is used as an alternative marker of vitamin D transcriptional activity, as it is a ligand-induced VDR-mediated target gene in intestinal cells (Pavek *et al*, 2009).

In primary human normal colonic mucosa (n=115), expression of *CDH1* was found to be associated with both *VDR* and *CYP3A4* (Table 9.5). As a total of 11 probes were tested in this analysis, the significance threshold was set at 0.0045 to correct for multiple testing. Although one of the two probes for VDR did not survive multiple testing, the probe that did (ILMN_2319952) was very highly significant *p*=5.44E-15) and explained a remarkable 51% of the variance in *CHD1* expression. As there are 7 reported protein-coding transcripts for *VDR*, of which 5 are poorly-supported transcript models according to the Ensembl genome database (URL9.1), it is reassuring that the ILMN_2319952 probe captures the two well supported protein-coding transcript models VDR-002 and VDR-004. On the other hand, expression of *FOXO1*, *FOXO3* (two out of three probes) and *FOXO4* are also very significantly associated with the expression of *CDH1* (Table 9.5), of which *FOXO4* appears to be the most significantly associated gene (*p*=1.62E-13).

Taken altogether, these results lend support to the hypothesis that the expression of *CDH1* is regulated by VDR and FOXO proteins. However, these associations are only suggestive of a mechanistic link, and do not shed light on the nature of this relationship and the direction of regulation, if any.

Illumina probe ID	**Gene**	**Spearman correlation with CDH1**		**Linear model adjusted for age, gender and sampling site**	
		rho	***p*-value**	**Estimate**	***p*-value**
ILMN_2319952	*VDR*	0.714	< 2.2e-16***	1.025	5.44E-15***
ILMN_1666203	*VDR*	0.235	0.011	1.282	0.015
ILMN_1772206	*CYP3A4*	0.375	3.73E-05***	1.023	3.96E-05***
ILMN_1738816	*FOXO1*	0.381	2.67E-05***	0.648	1.29E-05***
ILMN_1681703	*FOXO3*	0.528	1.34E-09***	0.578	1.23E-11***
ILMN_1712515	*FOXO3*	0.213	0.023	0.947	0.063
ILMN_1844692	*FOXO3*	0.429	1.67E-06***	0.558	3.30E-09***
ILMN_1712095	*FOXO4*	0.578	1.38E-11***	0.734	1.62E-13***
ILMN_3307977	*FOXO6*	-0.074	0.432	-0.516	0.255
ILMN_3311135	*FOXO6*	-0.098	0.299	0.054	0.907
ILMN_3311155	*FOXO6*	-0.008	0.936	-0.139	0.804

Table 9.5 The association between expression of *CDH1* and the expression of implicated genes in the human normal colorectal mucosa.

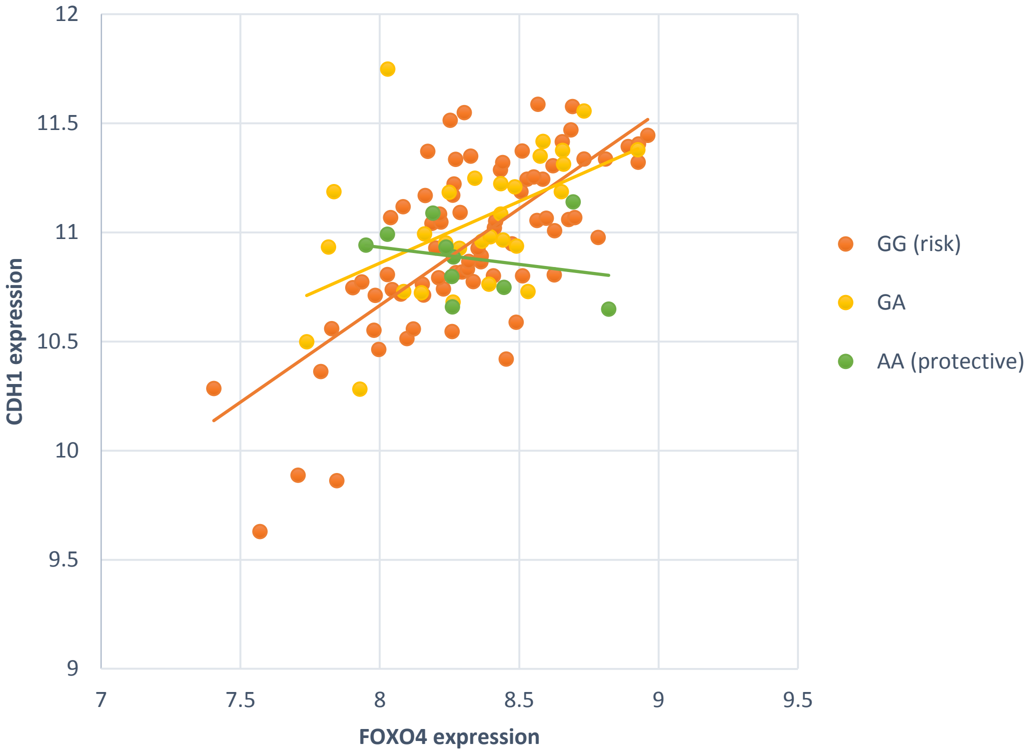
9.3.2 CRC risk variant rs9929218 modifies the association between *FOXO4* and *CDH1*

To gain insight into the direction of regulation, a two-way interaction analysis was performed with the *CDH1* variant, rs9929218. A statistically significant two-way interaction (p=0.0057) was observed between rs9929218 and *FOXO4* expression on the expression of *CDH1* (Table 9.6), suggesting that *FOXO4* influences the expression of *CDH1* and not vice-versa. Interaction analysis with rs9929218 was also carried out for *VDR* and the other *FOXO* family members, but no other further interactions were observed to be present.

The negative estimate of the interaction term between the main variables *FOXO4* and rs9929218 implies that there is negative synergy between them, i.e. their presence at the same time dampens their effect on *CDH1*. To illustrate this, the association between *FOXO4* and *CDH1* expression was analysed separately for each rs9929218 genotype (GG, GA and AA) (Figure 9.3). The relationship between *FOXO4* and *CDH1* appears to be modified by the rs9929218 genotype, where the gradient of the positive linear relationship between *FOXO4* and *CDH1* expression decreased with the number of A alleles. Although there is the possibility of a false positive result due to the small numbers of the AA genotype, this statistical interaction makes biological sense given that rs9929218 has been predicted to modify FOXO binding affinity, and offers a plausible mechanism for the plasma 25OHD-rs9929218 gene-environment interaction on CRC risk.

	Estimate	***p*-value**
Age	0.003	0.15
Gender	-0.085	0.11
Sampling site	-0.079	0.19
FOXO4	0.878	1.77E-14***
rs9929218	3.383	0.0053**
FOXO4**rs9929218**	-0.404	0.0057**

Table 9.6 The multivariate linear regression modelling for expression of *CDH1* in the normal mucosa demonstrating a significant two-way interaction between *FOXO4* and rs9929218.



rs9929218	N	Estimate	*p*-value
GG	76	0.863	4.05E-12***
GA	29	0.557	6.51E-03**
AA	10	-0.178	0.53

Figure 9.3 The association between expression of *CDH1* and *FOXO4* in the normal mucosa, analysed separately for each rs9929218 genotype (the major allele G is associated with a higher risk of colorectal cancer whereas the minor allele A has a protective effect).

9.3.3 *VDR* expression and a *VDR* polymorphism independently modifies the association between *FOXO4* and *CDH1*

A further analysis was carried out to examine the possibility of *VDR* influencing the relationship between *FOXO* and *CDH1* expression. A statistically significant two-way interaction (*p*=0.00617) was observed between *VDR* and *FOXO4* expression as determinants of *CDH1* expression (Table 9.7). The association between *FOXO4* and *CDH1* expression was analysed separately for each tertile of VDR expression (Figure 9.4). The relationship between *FOXO4* and *CDH1* appears to be modified by levels of *VDR*, where the influence of *FOXO4* on *CDH1* appears to decrease as *VDR* levels increased.

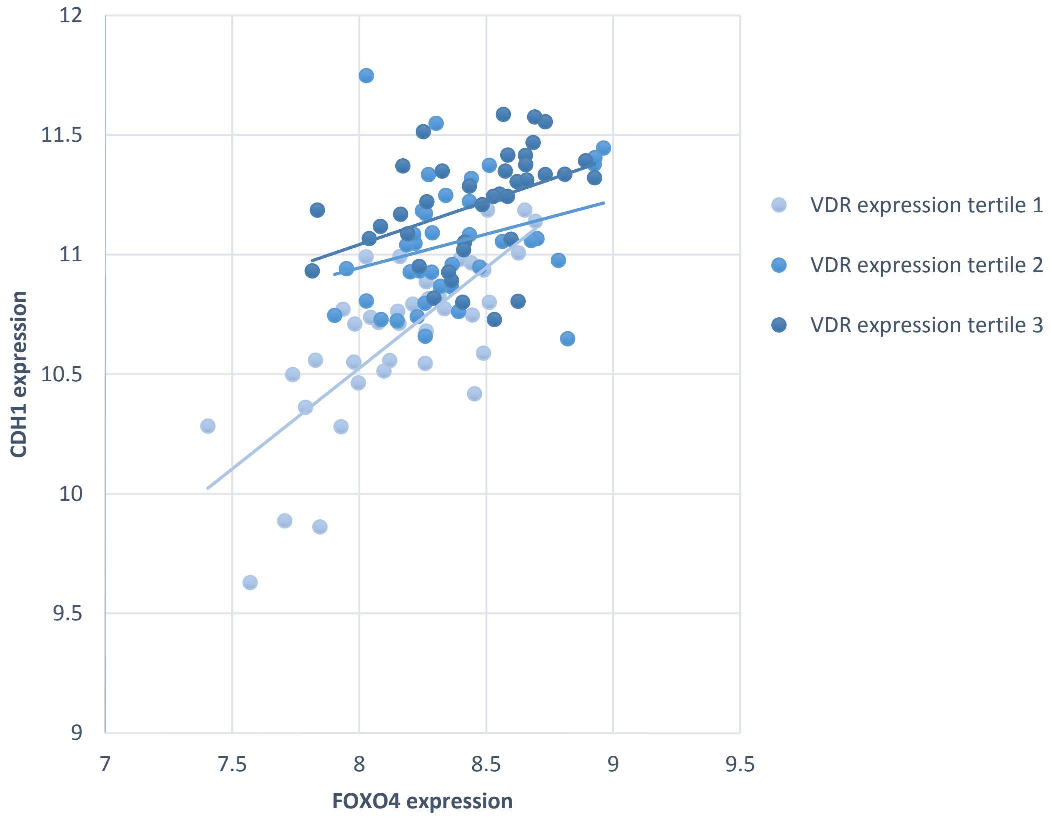
To find further supporting evidence for the interaction between *VDR* and *FOXO4,* two-way interaction analyses was carried out between *FOXO4* and *VDR* polymorphisms that have been reported to have a bearing on *VDR* function (as reviewed by Uitterlinden *et al,* 2004) and the risk of colorectal adenomas and cancer (Touvier *et al*, 2011; Bai *et al*, 2012) (Table 9.8). Interestingly, the *Fok*I polymorphism that is located in the start codon showed a significant two-way interaction with *FOXO4* (*p*=0.0076), where the influence of *FOXO4* expression on *CDH1* expression increased with the number of the major allele G (Figure 9.5). The major allele produces a protein that is shorter by three amino acids (Whitfield *et al*, 2001), and could conceivably influence the modulatory effect of VDR on FoxO4.

The independent effect of *VDR* expression and a *VDR* polymorphism on the *FOXO4-CDH1* association is strongly supportive of a direct association between VDR and FoxO4 that enhances FoxO4 binding to regulatory elements and the consequent transcription of *CDH1*.

In view of these significant two-way interactions implicating a functional relationship between *VDR*, *FOXO4* and rs9929218 on *CDH1* expression, a three-way interaction analysis was performed for these three variables. No significant three-way interaction was found (*p*=0.644) but the power of detection may be limited by the relatively small number of samples.

	Estimate	***p*-value**
Age	0.003	0.11
Gender	-0.070	0.11
Sampling site	-0.024	0.63
FOXO4	6.912	3.29E-03**
VDR	7.800	2.60E-03**
***FOXO4**VDR**	-0.848	6.17E-03**

Table 9.7 The multivariate linear regression modelling for expression of *CDH1* in the normal mucosa demonstrating a significant two-way interaction between *FOXO4* and *VDR*.

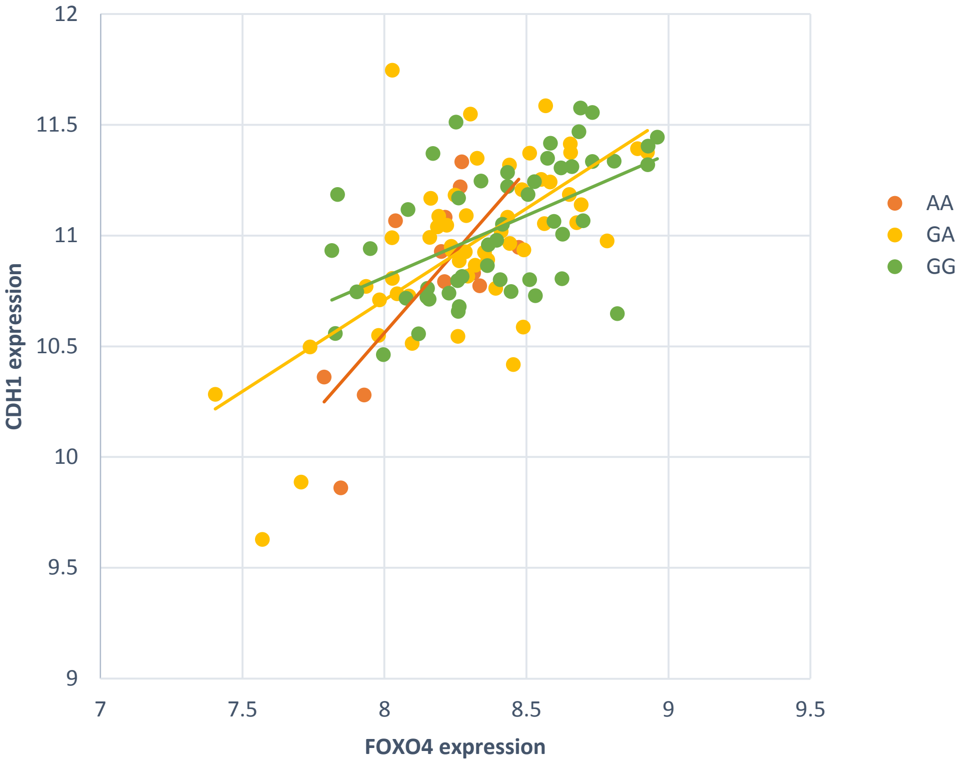


VDR expression	N	Estimate	*p*-value
Tertile 1 (7.0 – 7.49)	38	0.838	2.06E-07***
Tertile 2 (7.49 – 7.67)	38	0.205	0.178
Tertile 3 (7.67 – 8.25)	39	0.330	2.01E-02*

Figure 9.4 The association between expression of *CDH1* and *FOXO4* in the normal mucosa, analysed separately for each *VDR* expression tertile.

***VDR* polymorphisms**	**MAF**	**Estimate**	***p*-value**
FOXO4*_Apa_I (rs7975232)	0.46	-0.218	0.11
FOXO4*_Fok_I (rs10735810)	0.40	0.393	7.60E-03**
FOXO4*_BSM_I (rs1544410)	0.33	0.108	0.39

Table 9.8 Two-way interaction analyses between *VDR* polymorphisms and *FOXO4* expression in the linear regression modelling of *CDH1* expression. Model adjusted for age, gender and site of sampling.



FokI genotype	N	Estimate	*p*-value
AA	12	1.541	0.025*
GA	53	0.836	1.16E-07***
GG	50	0.487	3.90E-04***

Figure 9.5 The association between expression of *CDH1* and *FOXO4* in the normal mucosa, analysed separately for each rs9929218 genotype.

9.3.4 Tissue-specific effects of *VDR*, *FOXO4* and rs9929218 on *CDH1* expression

In a subset of matched PBMC (n=59), *CDH1* is not detected on the HT12 microarray and is only detectable at low levels by qRT-PCR. This is not surprisingly as it is well-established that E-cadherin is an adherens junction protein predominantly expressed in epithelial cells. However, recent studies have uncovered a role for this adhesion molecule in mononuclear phagocyte functions, where it regulates the maturation and migration of Langerhans cells, as well as the interaction between various immune cells and dendritic cells (as reviewed by Van den Bossche *et al*, 2013). Hence, it was thought to be of interest to investigate whether the genes and two-way interactions associated with *CDH1* expression are also present in PBMCs.

Similar to that in the normal mucosa, rs9929218 is not significantly associated with the expression of *CDH1* (*p*=0.67). The expression of *CDH1* is also not associated with the *FOXO* family members (*p*>0.43) or *VDR* (*p*=0.066). The two-way interactions between *FOXO4* and rs9929218, *VDR* and *Fok*I individually were not present in PBMC. This suggests that the postulated effect of *VDR* modulating the activity of *FOXO4* on cis-regulatory elements of *CDH1* is specific to the normal colorectal mucosa.

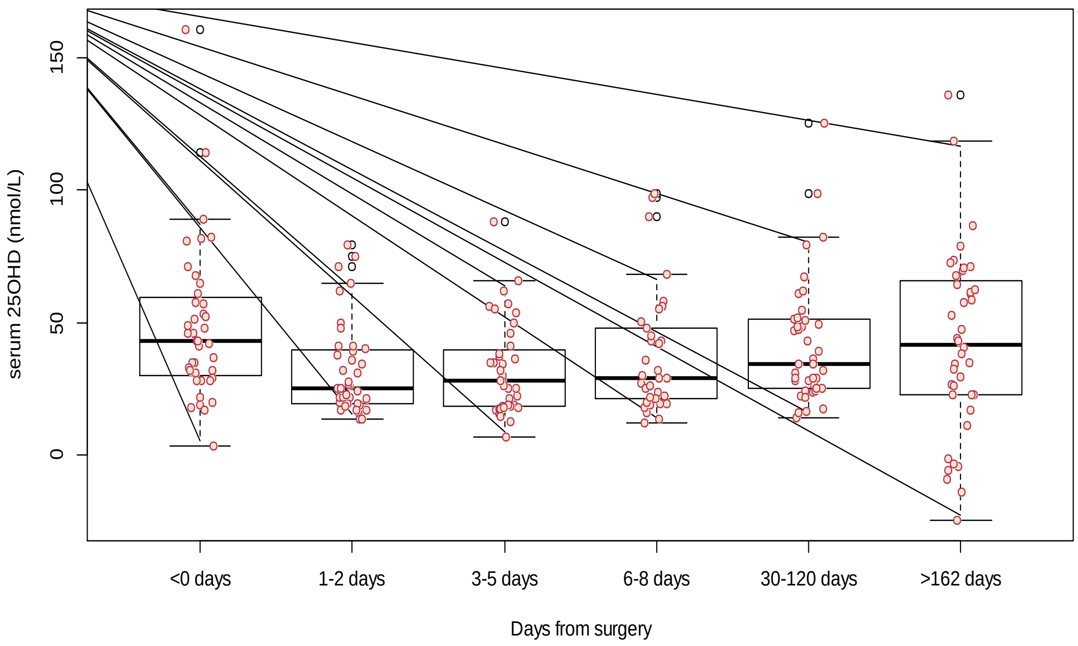
9.3.5 Analysis of the effect of serum 25OHD on normal mucosa *CDH1* expression

In view of the gene-environment interaction of plasma 25-OHD and rs9929218 on colorectal cancer risk, matched serum 25-OHD was retrospectively collected for a subset (n=83) of the normal mucosa used for the above analysis. Serum 25-OHD was not significantly associated with *CDH1* expression, and neither was there a statistically significant interaction between serum 25-OHD and rs9929218, *VDR, Fok*I or *FOXO4* on the expression of *CDH1* in the normal mucosa or PBMC. This is not surprising, as serum 25-OHD was collected at variable time points post-operatively, and may not accurately represent the intracellular vitamin D status of the normal mucosa tissue collected during the surgical procedure. It has been reported that that circulating 25-OHD is affected in patients undergoing cardiopulmonary bypass (Krishnan *et al*, 2010) and elective knee arthroplasty (Reid *et al*, 2011). Krishnan *et al* reported that plasma 25-OHD returned to baseline pre-operative levels by post-operative day 5, whereas Reid *et al* observed that 25-OHD remained significantly lower at 3 months post-operatively. Hence, it is possible that any association between vitamin D and *CDH1* expression may have eluded detection due to the changes in circulating 25-OHD that accompanies major surgery.

9.3.6 Serial sampling of circulating 25-OHD in patients undergoing large bowel surgery

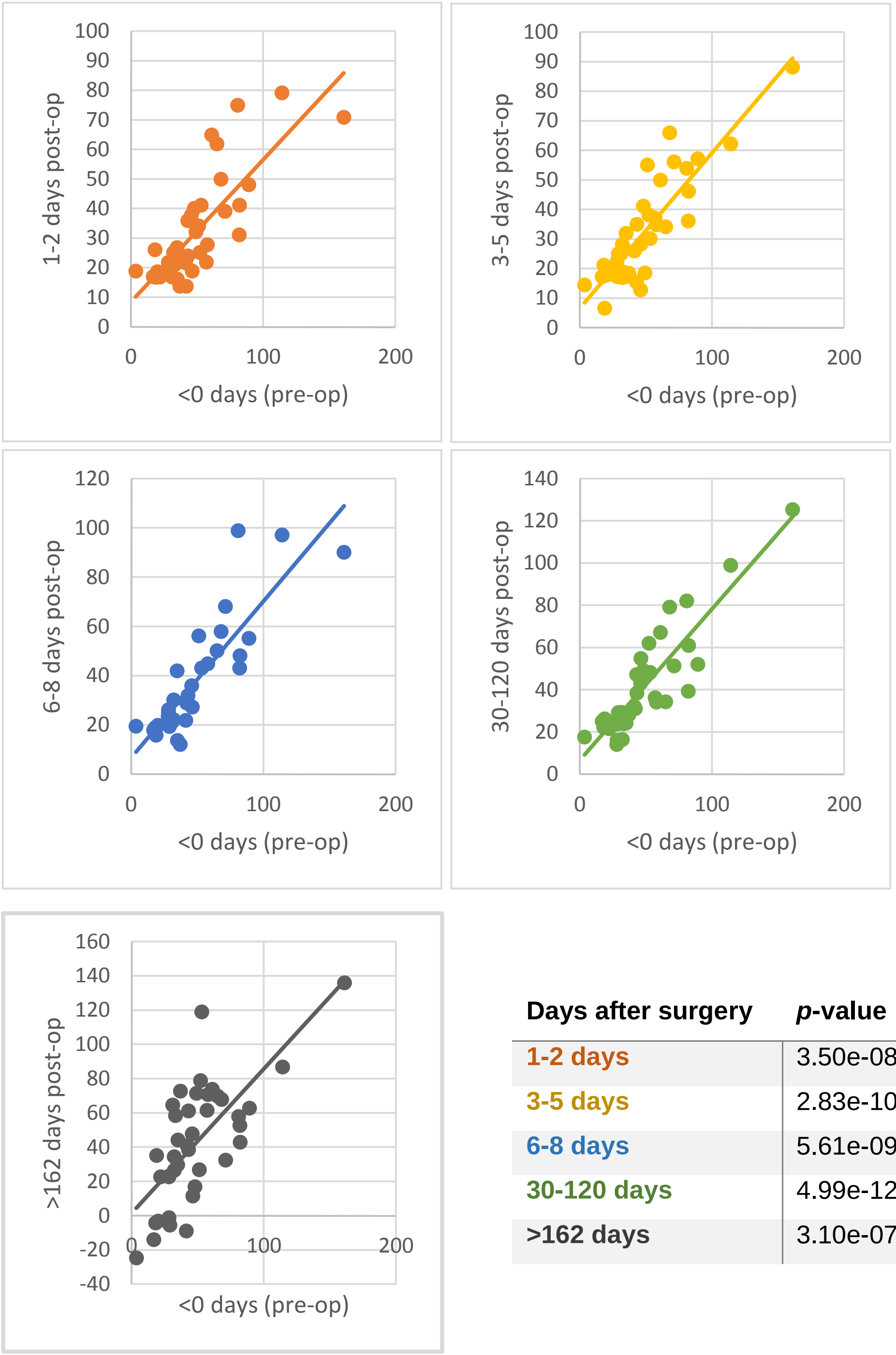
To address the fluctuations of circulating 25-OHD that may result from large bowel surgery, serial samples of serum were prospectively collected from patients undergoing elective large bowel resections for colorectal cancer (n=40) that consisted of one pre-operative sample and five post-operative samples. There was a significant reduction in circulating 25-OHD levels at the first four post-operative time points (Figure 9.6). The reduction was observed to be largest at 1-2 days post-op, with a diminishing effect as the number of days from surgery elapsed. At the final time-point (>162 days from surgery), the reduction was no longer significant, suggesting that circulating 25-OHD returned to baseline levels after approximately 5.5 months.

Interestingly, for each individual person, the levels of 25-OHD at every post-operative time point was very significantly associated with their pre-operative baseline levels (Figure 9.7). This indicates that although an operation impacted on patients' absolute 25-OHD levels, it does not influence their relative or ranked 25-OHD levels.



Days after surgery	<0 days (pre-op)	1-2 days	3-5 days	6-8 days	30-120 days	>162 days
Median (range)	43.0 (3.6-160.9)	25.2 (13.8-19.2)	28.2 (6.6-88.2)	29.0 (12.1-99.0)	34.2 (14.2-125.3)	41.9 (-24.8-135.9)
N	40	40	32	32	40	40
***p*-value**		8.38e-07***	8.64e-09***	2.32e-04***	8.26e-03**	0.075

Figure 9.6 Boxplots demonstrating serial circulating 25-OHD levels in patients undergoing large bowel resections for colorectal cancer. Each post-operative time point was compared to the pre-operative time point individually using the Wilcoxon signed-rank test.



Days after surgery	*p*-value
1-2 days	3.50e-08***
3-5 days	2.83e-10***
6-8 days	5.61e-09***
30-120 days	4.99e-12***
>162 days	3.10e-07***

Figure 9.7 Scatterplots demonstrating the association between pre-op and post-op 25-OHD levels. Each post-op time point was analysed separately using a linear regression model adjusted for age, gender, AJCC stage and type of surgery (open or laparoscopic).

9.3.7 Vitamin D treatment of human cell lines and colonic organoids

The two-way interactions influencing *CDH1* expression in the normal mucosa implicates the vitamin D signalling pathway in the regulation of *CDH1*, and suggests that the rs9929218 SNP modifies this regulation by altering the binding of *VDR*-interacting transcription factor *FOXO4*. To experimentally investigate these observations, six colorectal cancer cell lines that are homozygotes for the rs9929218 polymorphism (AA=3, GG=3) were selected for treatment with calcitriol (1α,25-dihydroxyvitamin D), which is the active metabolite of vitamin D. It was hypothesised that firstly, *CDH1* expression would be induced by calcitriol, and secondly, this response will vary according to the rs9929218 genotype.

Baseline *VDR*, *CDH1* and *CYP24A1* expression were first checked for each of these cell lines (Figure 9.8). Triplicate time courses were carried out for each cell line, and the *CDH1* response was calculated by the fold change from controls treated with the ethanol carrier at each time point (Figure 9.9). *CYP24A1*, a well-known target gene of *VDR*, was measured as a positive control for calcitriol-dependent transcriptional response (Figure 9.9). Overall, the induction of *CYP24A1* appears to peak at 16-24 hours, with the exception of COLO205 that had a relatively small fold change. Baseline levels appeared to influence the magnitude of the calcitriol-induced fold change i.e. cell lines with higher baseline levels of *CYP24A1* demonstrated smaller fold changes.

Of the six cell lines, SW480 showed the most convincing response in *CDH1*- a six-fold change was observed after 16 hours of calcitriol treatment and this persisted up to 48 hours (Figure 9.10). The response after 24 and 48 hours was variable between the replicates and hence not statistically significant, but a minimum of a four-fold change was still present. This large differential response may be partly due to the fact that baseline *CDH1* levels are very low in SW480 (Figure 9.8) and the any absolute increase in *CDH1* levels will be reflected as a large fold change. Two other cell lines, LS174T and SW48 also showed a significant induction of *CDH1* at 16 hours of treatment, albeit of a lesser magnitude (>two-fold change). There was no obvious influence of the rs9929218 genotype on the *CDH1* response. However, there were large differences in baseline levels of *CDH1* and *VDR*, both of which could in theory influence the effect of calcitriol stimulation on *CDH1* expression fold change, hence displacing any effect of rs9929218. Nevertheless, the induction of *CDH1* expression in three out of six colorectal epithelial cell lines supports a link between vitamin D and the regulation of *CDH1* expression in the normal mucosa.

As discussed in previous chapters, the use of cell lines for assessing SNP function is challenging and suboptimal at best due to the cellular genomic and karyotypic abnormalities that they have accumulated. To address this limitation, the culture of colonic organoids derived from primary human large bowel epithelium was instigated as a non-aberrant in-vitro model system (Figure 9.11). It has to be noted that the organotypic culture used in this thesis is at its preliminary stages, and the methods are still in need of optimisation for the organoids to proliferate in a robust and reproducible manner. Presently, these organoids do not survive for more than 10 days, hence, they were treated at day 5 for a comparison of *CDH1* expression in response to calcitriol treatment. Based on the time course studies in cell lines, the organoids were treated for 16 hours to elicit a maximal response. A baseline level of gene expression was also quantified at day 0 for comparison.

Unfortunately, there were recurring issues with fungal infections of the organoid culture, and only three of the five organoid cultures survived until calcitriol treatment at day 5. Hence, no meaningful statistical analyses were able to be performed. Two of these organoid cultures were of the rs9929218 (GG) genotype, whereas one was of the rs9929218 (AA) genotype (Figure 9.12). Firstly, it can be observed that the two positive control genes, *CYP24A1* and *CYP3A4*, have very low baseline levels. In fact, *CYP24A1* was undetectable in two out of the three samples, which concurs with the expression data of the normal mucosa. Treatment with calcitriol induced a response in all three organoid cultures, which is reassuring of calcitriol penetration through the Matrigel scaffold and a vitamin D-dependent transcriptional response. Baseline levels of *CDH1* and *VDR* appears to be closely matched, which again is in concordance with the normal mucosa expression where these two genes are very significantly correlated. Calcitriol treatment appears to induce a 1.5 fold increase in *VDR* expression in one of the organoid cultures, but this was not observable in the other two. Rather interestingly, calcitriol treatment appears to induce a small increase in *CDH1* in the organoids with the rs9929218-AA genotype (n=1), which was not seen in organoids with the rs9929218-GG genotype (n=2). However, this sample had higher baseline *CDH1* levels at day 0, which had reduced by almost half at the point of calcitriol treatment at day 5

These preliminary results suggest that normal mucosa derived-organoids could serve as an effective model system to demonstrate common allele-specific effects that are only apparent under cellular perturbations such as a ligand-dependent regulation, but will require replication before a suggestion of an allele-specific differential effect can be convincingly made.

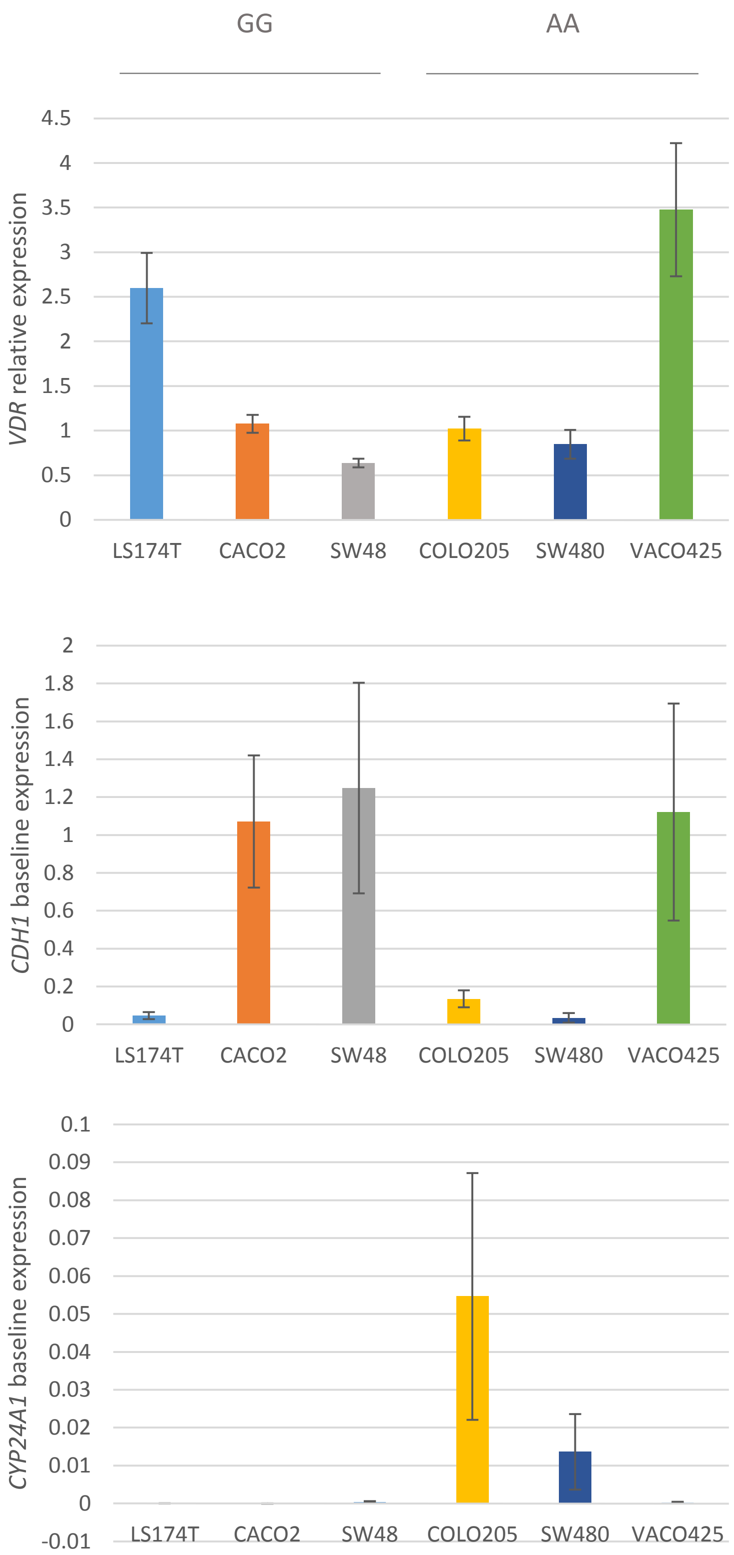


Figure 9.8 Baseline expression of *VDR*, *CDH1* and *CYP24A1* in cell lines selected for calcitriol treatment. Triplicate measurements were taken. Error bars=SEM.

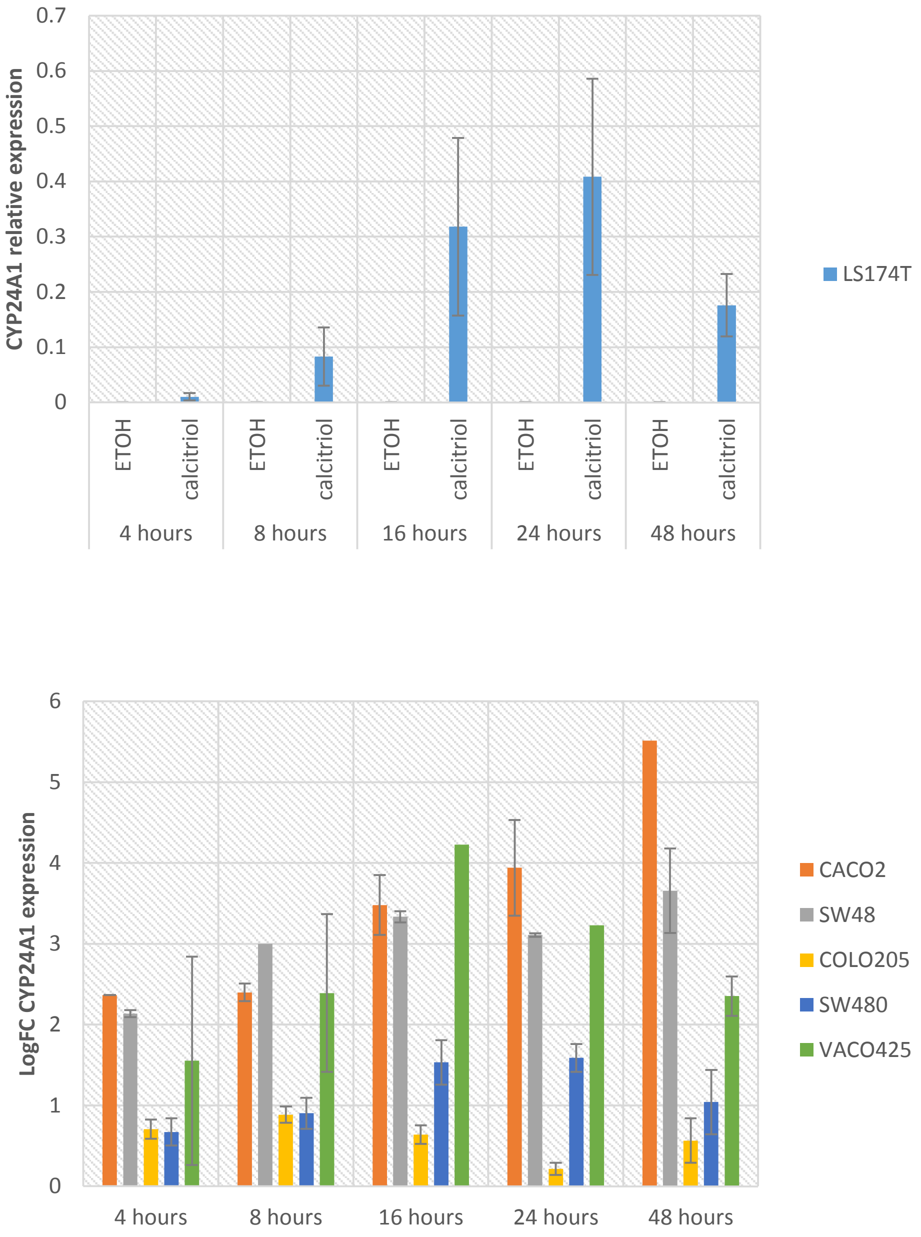


Figure 9.9 *CYP24A1* induction by calcitriol treatment for a series of time points in six colorectal cancer cell lines. A) Relative expression of *CYP24A1* in LS174T is presented individually for the ethanol control and calcitriol treatment as fold changes cannot be calculated from an undetectable baseline. B) Log fold change of *CYP24A1* expression between calcitriol treatment and ethanol control in the other five tested cell lines. Error bars are not presented when fold changes are calculable for only one replicate due to undetectable baseline levels.

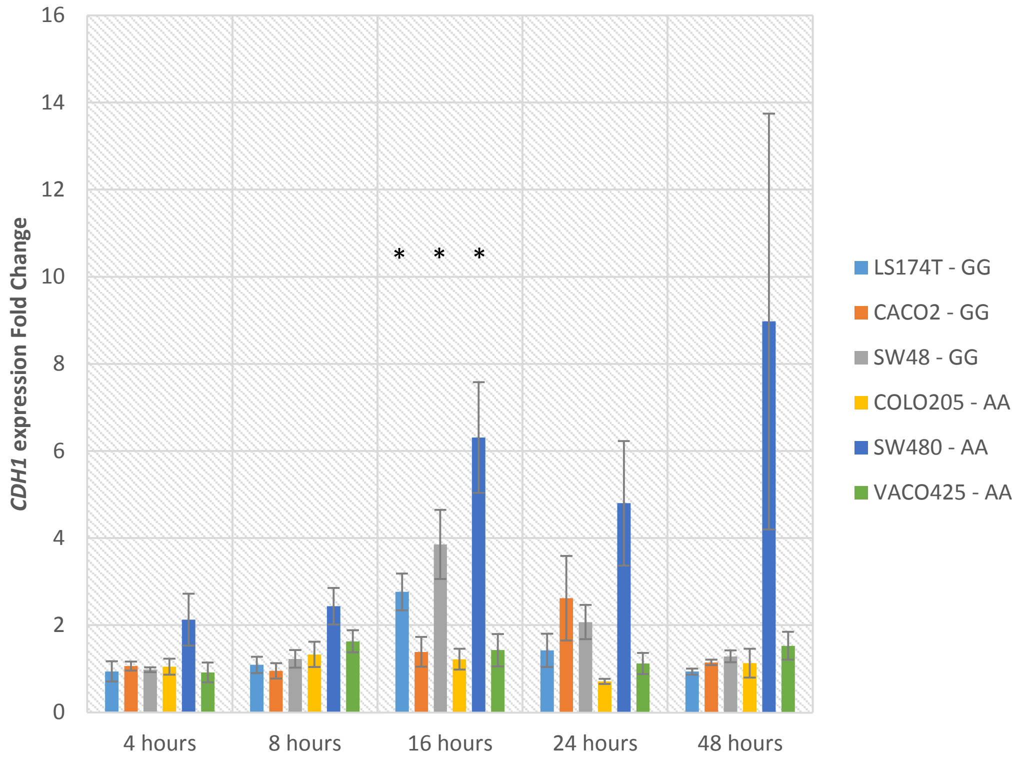


Figure 9.10 *CDH1* expression fold change with calcitriol treatment for a series of time points in six colorectal cancer cell lines, performed in triplicates. Error bars=SEM. Student unpaired t-tests were performed on the fold change compared to the ethanol-treated controls, * indicates p<0.05.

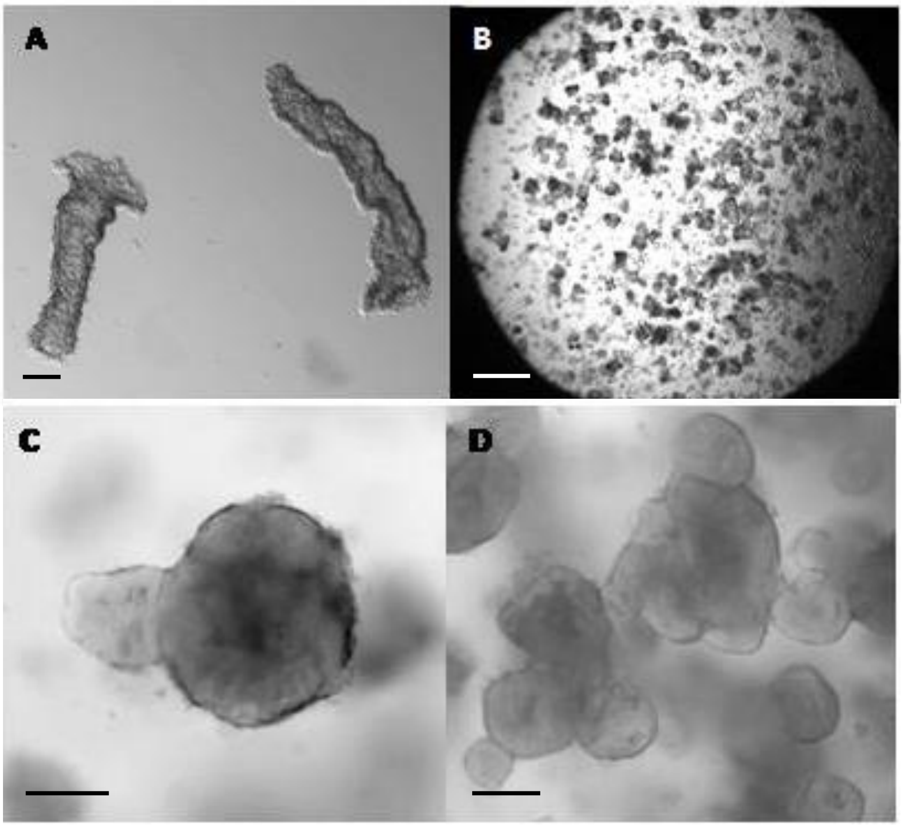


Figure 9.11 Human normal colon organoids. A) Colon crypts disaggregated from surrounding stroma. B) Crypts disrupted and seeded. C) Crypt like structures budding at day 5. D) Organoids at day 7 with multiple crypt buds. *Scale bar*: A, C and D - 50 µm; B - 250 µm.

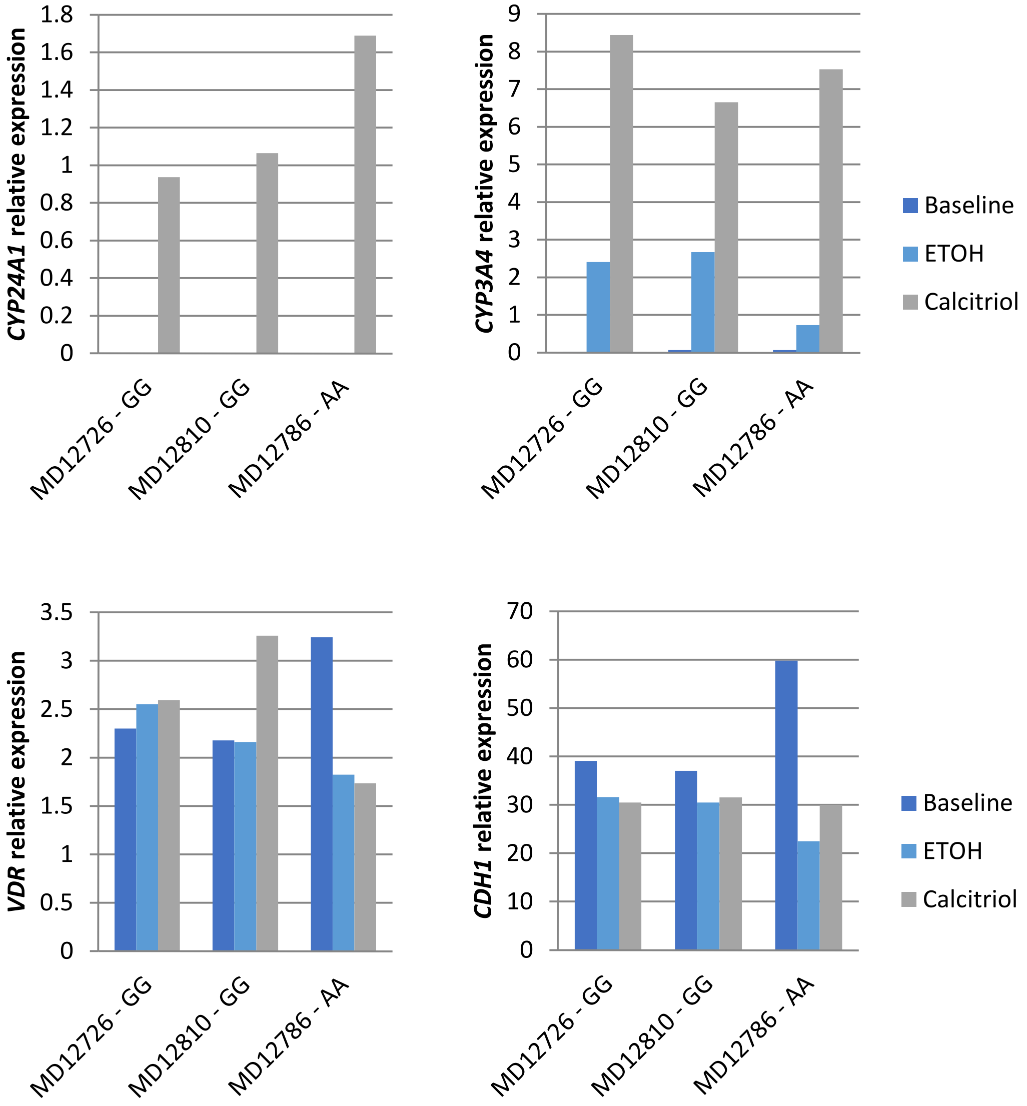


Figure 9.12 Relative expression of *CYP24A1*, *CYP3A4*, *VDR* and *CDH1* in human colonic organoids derived from the normal mucosa of three patients. Each MD number represents crypts derived from an individual patient, with their rs9929218 genotype indicated. A baseline measurement was taken at day 0 after the colonic crypts were dissociated from the stroma. Organoids were treated at day 5 in culture with 1% ETOH (negative control) or 100nM calcitriol for 16 hours. Relative expression was measured by qRT-PCR, using *ACTB* as a reference gene.

9.4 Discussion

The chemopreventative role of vitamin D in colorectal cancer has been the focus of many recent studies. Various approaches have been used to estimate vitamin D status, including direct measures of circulating 25-OHD, surrogates or determinants of vitamin D such as dietary intake and sun exposure estimates (as reviewed by Giovannucci, 2010). Although confounding factors cannot be entirely excluded, the consistency of these associations with CRC are highly suggestive of a causal association. There have been two randomised controlled trials investigating the effect of vitamin D supplementation on colorectal cancer risk, both of which failed to demonstrate an effect on CRC incidence (Trivedi *et al*, 2003; Ding *et al*, 2008). However, these studies are limited by the small number of participants, low doses of vitamin D and inadequate trial duration to demonstrate an effect. Hence, the gene-environment interaction between 25-OHD and a known CRC susceptibility variant at the *CDH1* gene (rs9929218) (Zgaga *et al*, unpublished) is of significant interest, as not only does it lend weight to the suggestion of causality, it may also explain a proportion of the missing heritability of colorectal cancer. It is critical to identify the molecular mechanism underlying this relationship, as it will provide validation to the epidemiological findings and consequently inform the rational development of targeted preventative and therapeutic strategies. Given that the role of *CDH1* (E-cadherin) in cancer is well-established, it was hypothesised that this gene-environment interaction manifests its effects on the regulation of its transcription, particularly as its expression has previously been shown to correlate with *VDR* expression in colorectal tumours (Pena *et al*, 2005).

The correlation of *CDH1* expression in the normal colonic mucosa with *VDR* and *CYP3A4,* an intestinal *VDR* target gene, is suggestive of *CDH1* regulation by vitamin D activity in the large bowel epithelium. The strong correlation of *CDH1* expression with the *FOXO* transcription factors also implicates a role for them in the regulation of *CDH1*. The regulatory relationship suggested by these correlations have been previously demonstrated in-vitro using various cell line models – calcitriol-dependent *VDR* regulation of *CDH1* in the colorectal cancer cell line SW480 (Palmer *et al*, 2003), and *FOXO*-mediated regulation of *CDH1* in urothelial cells (Shiota *et* *al*, 2010) and kidney epithelial cells (Carew et a, 2011). A recent study demonstrated that ligand-bound VDR induce the dephosphorylation and activation of FoxO proteins to regulate common VDR/FoxO target genes in a squamous cell carcinoma cell line (An *et al*, 2010). Hence, the finding of multiple two-way interactions between *FOXO4* levels-rs9929218, *FOXO4* levels-*VDR* levels and *FOXO4* levels-*FokI* is exciting, as it alludes to a biologically plausible in-vivo co-regulatory relationship between *VDR* and *FOXO4* on the expression of *CDH1* that is modified by an established risk SNP, and provides support to the hypothesis driven by the epidemiological and bioinformatics analysis. However, the number of samples in this study is very small and must be validated independently in larger studies.

Circulating 25-OHD is the most frequently used biomarker of vitamin D status in clinical settings and epidemiological studies, as it accounts for both endogenous synthesis in the skin and vitamin D intake, and has been shown to vary widely in human populations (as reviewed by Jacobs *et al*, 2011). Although there is a lack of association between *CDH1* expression and circulating 25-OHD (singularly or as a factor in two-way interaction analysis), it should be recognised that there are limitations inherent to the use of 25-OHD in associative studies, which is especially pertinent to the study presented in this chapter. Firstly, circulating 25-OHD may not effectively capture intracellular vitamin D status due to the dynamics and variability in local tissue-level conversion of 25-OHD to the active metabolite 1,25(OH)$_2$D, and secondly, the sampling of serum 25-OHD was carried out post-operatively at varying time points from the treatment for cancer. As shown in the serial samples of 25-OHD of patients undergoing large bowel surgery, the procedure impacts on circulating 25-OHD levels with effects lasting well beyond the hospital stay (approximately a week). This could be due to a combination of various factors such as re-distribution of vitamin D metabolites induced by a general anaesthetic, IV fluids, inflammatory responses, and the lack of sun exposure during the rehabilitation period. For the normal mucosa gene expression study, patients undergoing adjuvant chemotherapy were not excluded and it is also possible that this may also have an influence on 25-OHD levels. Hence, it is very likely that 25-OHD levels used in the gene expression correlation study do not accurately reflect the vitamin D status of the tissue harvested from the surgically resected large bowel specimens. Interestingly, the results from the serial samples of 25-OHD indicate that the impact of surgery on *absolute* 25-OHD levels did not affect the *relative* levels of 25-OHD at all the time points examined, suggesting that sampling at a consistent time from the operation could be an acceptable substitute for pre-operative sampling. In other words, the time from the operation to sampling could be included as a variable in the regression model to account for the effects of surgery, which is potentially important in association studies of cancers where surgery is the mainstay of treatment and patients are recruited at varying time periods after surgical treatment. Future work should include further assessment of the serum 25-OHD serial sampling data using repeated measures analysis, as this would remove between-subjects variability and could improve the power of the test in detecting significant differences between means.

The calcitriol treatment of CRC cell lines and human colorectal organoids provide preliminary evidence that *CDH1* is up-regulated by vitamin D activity. Going forward, more detailed functional studies are now required to robustly elucidate the mechanism of this regulation, as well as any allele-specific effect of rs9929218. Using cell lines that have shown a *CDH1* transcriptional response to calcitriol treatment, immunoprecipitation and western blotting will be able to demonstrate any post-translational modifications of the FoxO proteins that are mediated by VDR, ChIP will reveal DNA-protein binding at the rs9929218 locus, and gene depletion of *VDR* and *FOXO* will establish the role of these transcription factors on *CDH1* transcription. EMSAs and luciferase reporter plasmids are also useful in-vitro assays that can show an allele-specific effect of rs9929218. Replication with the human organoid culture will consolidate the functional relevance in the non-transformed tissue state, but this may not always be technically feasible due to the limited amount of tissue available for the set-up of culture and the large amounts of cellular extract required for some of these assays. Hence, it is of interest to optimise the viability of the colonic crypts and the growth conditions for allow adequate expansion of the culture for parallel functional experiments.

In summary, the data presented in this chapter has demonstrated an approach using gene expression data derived from colorectal primary tissue, cell lines and human organoids to gain insight into the molecular mechanisms underlying a gene-environment interaction involving a common CRC susceptibility variant and vitamin D status. Although the evidence is largely observational and preliminary, it suggests that there is scope for further discovery and sets the groundwork on which further functional studies can be built. There is enormous potential and value in pursuing the molecular mechanism underlying this gene-environment interaction as the level of vitamin D is modifiable with supplementation, which has a relatively safe side-effect profile. Informed, appropriate selection of those that would benefit most from an improvement in vitamin D levels can potentially lead to a large impact on the prevention of colorectal cancer.

Chapter 10

Summary and Discussion

10.1 Summary

In the last decade, the use of GWAS on large, well-characterised case-control cohorts of colorectal cancer has facilitated the identification of greater than 25 common genetic variants that carry with them an increased predisposition to colorectal cancer. As the majority lie within non-coding regions, the underlying causal mechanism is to-date poorly understood for the majority of these loci. The work presented in this thesis has demonstrated that a number of these genetic variants also influence gene expression levels, strongly suggesting that they confer risk, at least in part, by modifying regulatory mechanisms.

The hypothesis that CRC-associated genetic variants influence gene expression was tested by two approaches - an agnostic approach that utilised eQTL analysis, and a hypothesis-driven approach that specifically examined the expression of target genes and regulatory pathways of an established risk locus. It was thought that these heritable influences on gene expression are likely to be subtle, hence there was a strong emphasis on the methodology and production of robust data to minimise experimentally-induced non-biological variations and consequent erroneous conclusions. Chapter 3 described the development of a reproducible protocol that ensures the extraction of high-quality RNA from primary colorectal mucosal tissue for reliable gene expression profiling, whereas Chapter 4 focused the selection and validation of context-specific reference genes for qRT-PCR. The detection of differential expression profiles in relation to clinicopathological features (Chapter 5) allowed internal validation of the sample and data processing, and also highlighted the importance of accounting for these potential confounders in the subsequent expression analysis.

The systematic analysis of the association between 25 established risk loci and expression of *cis*-genes (Chapter 6) provides evidence to support the hypothesis that these risk loci exert their effects on CRC risk by having tissue-specific eQTL effects on gene expression. Expression fine-mapping of these eQTL associations identifies putative functional variants, some of which are also better at predicting CRC risk, thus making them likely to be the causative variants. Chapter 7 follows up the Xp22.2 eQTL/risk locus with functional assays, validating the experimental rationale of *cis*-eQTL analysis and expression fine-mapping. By association, the target gene of this locus, *SHROOM2*, is a candidate in the predisposition to CRC. Little is known of *SHROOM2* in the context of CRC, and Chapter 8 outlines an investigative approach with preliminary functional data suggesting that *SHROOM2* has a possible role in cell cycle regulation and is likely to be expressed at the top of colonic epithelial crypts.

Although the eQTL analysis has produced risk loci-expression associations, the functional effects of many of the loci remain unexplained. Chapter 9 takes an alternative hypothesis-driven approach to understand the mechanism underlying the 16q22.1 locus, which has recently been shown in a gene-environment interaction analysis to modify the protective association of vitamin D levels on CRC risk. Variant-expression and expression-expression interaction analyses support a role for the vitamin D signalling pathway in the modulation of the heritable variation in *CDH1* expression, demonstrating a candidate-based approach in deciphering the link between genetic locus and CRC susceptibility.

10.2 A hypothesis-free discovery of candidate causal variants and genes: utility of tissue-specific eQTL analysis

It has been established that common genetic variants contribute to the risk of colorectal cancer, and the post-GWAS challenge is to elucidate how these risk variants specifically influence the development of colorectal cancer. By examining the expression of cis-genes neighbouring these CRC risk variants in normal colorectal mucosa and matching peripheral blood, I have demonstrated at least five local eQTL associations for these risk variants in each of the tissue types (Chapter 6), agreeing with the published observation that trait-associated SNPs are more likely to be eQTLs (Nicolae *et al*, 2010). The majority of these associations were tissue-specific; even when the risk variants were eQTLs in both tissue types, their target genes did not overlap, consistent with reports in the literature that disease-associated variants tend to exert more cell-type specificity (Fu *et al*, 2012; Brown *et al*, 2013). The presence of CRC-associated eQTL in colonic and extra-colonic tissue, as well as the tissue-specificity they exhibit, is interesting. Assuming that the cell-type that harbours the intermediate phenotype of transcript abundance contributes to the transformation of the cell of cancer origin, the extra-colonic eQTLs suggest that alterations in extra-colonic cells may indirectly modify CRC susceptibility in a non-cell autonomous fashion. This tissue-specificity also emphasizes the importance of selecting the relevant target tissue for the examination of eQTLs, particularly as previous examination of publically available LCL eQTL databases did not reveal convincing eQTL effects, with the exception of the 6p21.2 (Dunlop *et al*, 2012) and 18q21.1 (Broderick *et al*, 2007), where there was some evidence of association to the expression of neighbouring genes. In view of the heterogeneity of cell-types and transcriptional signatures within the intestinal epithelial crypts, risk eQTLs could in fact be specific to a particular crypt compartment or a particular epithelial cell-type, and it would be of interest to test this hypothesis with single-cell gene expression analysis.

The two best associations were seen in the colorectal mucosa, at 11q23.1 and Xp22.2, with adjusted association *p*-values in the order of 10e^{-09}, which fits in with the expectation that eQTL associations observed in the originating tissue giving rise to the tumour are likely to be more informative (Freedman *et al*, 2011). The target genes *COLCA1*, *COLCA2* and *SHROOM2* lie adjacent to the risk loci, and have not been previously known to be associated with cancer, possibly representing novel pathways/molecular networks that are involved in cancer initiation and progression. eQTL analysis of susceptibility loci in other tissue types such as the liver and prostate have been shown to be of value in identifying target genes that influence disease susceptibility (Musunuru *et al*, 2010; Pomerantz *et al*, 2010). The other three risk/eQTL loci identified in the colorectal mucosa were 12q13.12 (*SPATS2*), 8q23.3 (*UTP23*) and 1q41 (*HLX*), but their effects were much weaker and validation studies will be required.

In the PBMC, the eQTL effects of risk variants were also weak, but it should be noted that not only were the sample size considerably smaller, peripheral blood was obtained at inconsistent time points after the operation, increasing the variation ‘noise’ and reducing the power to detect eQTLs. Despite this, there were significant cis-associations even after multiple-testing correction, targeting genes that have been previously implicated in cancer biology such as *CERS5* (Ceramide synthase 5) and *LIMA1* **(LIM domain and actin-binding protein 1)**. Although eQTLs in PBMCs may not be as directly relevant to those detected in colorectal mucosa, they could be of interest as they may reflect indirect effects in immune cells that induce changes in the colorectal epithelium by modulating the stromal microenvironment, particularly as inflammatory processes are known to contribute to the development of CRC. Nevertheless, further study is required as it is unclear whether peripheral blood mononuclear cells are appropriate surrogates for mucosal immune cells. The association of the risk locus 20p12.3 with *RP11-19D2.2,* an uncharacterised long intervening non-coding RNA transcript is interesting in principle. Together with the evidence suggesting that *COLCA1* and *COLCA2* are also long non-coding RNAs (Smillie, pers. comm.), this lends support to the notion that low-abundance unannotated lncRNAs are transcribed from cancer risk loci and mediate risk by facilitating a wide repertoire of regulatory functions. Deep sequencing of transcripts derived from targeted regions with techniques such as RNA- CaptureSeq (Mercer *et al*, 2011) will allow targeted interrogation of different populations of RNA in relation to risk loci genotypes. More generally, RNA-Seq techniques have increased coverage over microarrays, providing the ability to look at alternative gene spliced transcripts, post-transcriptional modifications, gene fusion and allele specific expression. This would allow better definition of the transcriptome and ultimately be of greater value in detecting changes associated with risk alleles.

After the initial identification of eQTL associations, expression fine-mapping of these individual risk loci was performed using data from high-density genotyping arrays. For each eQTL loci, candidate functional variants for expression were compared for their effects on CRC risk, with the rationale that variants that better explain both target gene expression and CRC risk are more likely to be causal. Using this approach, candidate variants for the 11q23.1 (*COLCA1* and *COLCA2*), Xp22.2

(*SHROOM2*) and 12q13.12 (*CERS5*) loci were identified. Functional assays are required to hone in on the causal variant – this was demonstrated for the Xp22.2 locus in Chapter 7, where gene reporter assays showed marked differential transcriptional activity with Indel24 which was not seen with the tagging SNP, nor matched by the alternative candidate SNP rs5934685.

During the progress of this research, two reports of CRC risk variants exhibiting eQTL effects on cis-genes in colorectal tissue were published independently (Loo *et al*, 2012; Closa *et al*, 2014), however, not all the results were in agreement with my findings. Loo *et al* identified four genes (*ATP5C1*, *DLGAP5*, *NOL3* and *DDX28*) at three risk loci with differential expression levels as a function of genotype, none of which were nominally significant in my samples. Closa *et al*'s findings matched more closely to those of mine, implicating the 11q23.1, Xp22.2 and 12q13.12 loci as local eQTLs that similarly affected *COLCA1, COLCA2* and *SHROOM2* expression, as well as additional target genes *GPR143* and *DIP2B.* Although the independent replication of part of my findings is reassuring validation, the discrepancies in the others suggest that the eQTL analysis are subjected to errors induced by non-biological factors such as sample sizes and study power, genotyping and imputation methods, microarray platforms utilised and the sampling procedure (surgery or colonoscopy biopsies). Both studies also examined tumour tissue as well as normal tissue, which could have harboured large regulatory aberrations masking the subtle eQTL effects associated with germline variation. Closa *et al* also examined *trans*-eQTL effects of CRC risk loci on genome-wide gene expression, and found that two of the loci with *cis*-eQTL activity (11q23.1 and 12q13.12) also exhibited *trans*-associations with the expression of multiple genes, albeit weaker than the *cis*-associations. Although this is suggestive that the *trans*- associations are related to common transcriptional networks, this does not exclude the possibility that trans-eQTL activity could account for the yet unexplained function of other risk loci, as *trans*- associations tend to be indirect and hence weaker. Hence, it would still be of interest to examine for trans- associations in the normal colorectal mucosa, but a larger sample size may be required to increase the power of detection.

10.3 Identification of a novel causal variant and candidate cancer susceptibility gene at the Xp22.2 locus

To verify the eQTL associations identified at CRC risk loci, it is important that the causal variant is defined and the underlying regulatory mechanism delineated. The identification of the molecular mechanisms underlying the Xp22.2 eQTL locus is validation of the experimental rationale of examining expression as an intermediate phenotype. By using a combinatory approach of targeted resequencing and fine-mapping of the Xp22.2 risk locus with *SHROOM2* expression, two candidate functional variants (rs5934685 and the novel Indel24) were identified to be more strongly associated with both expression and risk than the tagging SNP rs5934683 (Chapter 7). Although conditional modelling of *SHROOM2* expression cannot exclude the possibility of independent association signals, conditional analysis of a case-control study supports Indel24 as the driver signal. Indeed, *in vitro* luciferase gene reporter assays indicates that the novel Indel24 is the most likely functional variant modulating regulatory control of transcription. *In silico* data from ENCODE ChIP-seq studies indicates that Indel24 resides within the binding sites of NF-YA and NF-YB, two transcription factors that bind cooperatively as two subunits of the trimeric NF-Y transcription factor complex. siRNA depletion of NF-YA and NF-YB singularly was associated with a reduction in transcription as observed by gene reporter activity, as well as endogenous *SHROOM2* levels in CRC and RPE cell lines, indicating that Indel24 may be modulating transcription of *SHROOM2* by altering the DNA-binding affinity of NF-Y. Indeed, on the minus strand, the NF-Y consensus binding motif CCAAT is present within the insertion allele of Indel24 with three other CCAAT motifs flanking Indel24. The introduction of mutations to the CCAAT motifs in a series of reporter constructs s provided further evidence that NF-Y is involved in the Indel24 modulation of differential transcriptional control, and that Indel24 modifies NF-Y binding affinities by altering the spacing between its functional binding motifs and not by donating an extra binding site. Aside from demonstrating the function of the causative eQTL variant at the Xp22.2 risk locus, this work also exemplifies how structural variation at non-coding regions can influence the activity of control elements.

The target gene of Xp22.2 eQTL, *SHROOM2*, is an interesting candidate for colorectal tumourigenesis, as it has previously been shown to have a role in cell morphogenesis during endothelial and epithelial tissue development (Lee *et al*, 2009; Farber *et al*, 2011), cytoskeletal organisation (Dietz *et al*, 2006), tight-junction stabilisation (Etournay *et al*, 2007) and cell contractility and migration (Farber *et al*, 2011), all of which are aspects often implicated in cancer biology. *SHROOM2* co-expression analysis in the normal colorectal mucosa is suggestive of a role in cell cycle regulation, which is also corroborated by the transcriptomic analysis of cell lines with siRNA depletion of *SHROOM2* (Chapter 8). Furthermore, the transcription factor implicated in *SHROOM2*'s transcriptional control, NF-Y, is known to activate the transcription of various cell cycle genes (Muller and Engeland, 2010), indirectly adding support to the postulation that *SHROOM2* exerts its tumour suppressive effects by influencing cell cycle progression. It would therefore be of considerable interest for future work to include experiments that would directly implicate *SHROOM2* in the regulation of cell cycle. The availability of compelling new tools recently reengineered within the unit, such as the bicistronic Fucci2a system and the R26Fucci2aR mouse model (Mort *et al*, 2014), provides an attractive collaborative opportunity with local expertise for the investigation of *SHROOM2*'s role in cell cycle dynamics, both in cell culture and during mouse embryonic development. In the planning and design of such functional assays, consideration should be given to the fact that aberrant gene activity of the inactive X-chromosome are often seen in neoplastic processes leading to perturbed dosage of X-linked factors. This is particularly relevant as some of these genes are known to be involved in cancer promotion and could confound functional phenotypes thought to be related to *SHROOM2*.

In the same vein as the considerations about the cell-type specificity of eQTLs, the localisation of *SHROOM2* expression within the colorectal epithelial crypt is of interest as it sheds light on protein function and its role in the development of cancer. As a specific antibody to *SHROOM2* was lacking, indirect evidence from *SHROOM2* co-expression analysis with cell-specific marker genes in the normal colorectal mucosa pointed towards expression in the crypt-top mature enterocytes, suggesting that *SHROOM2* may contribute to their tumour-initiating capacity via a 'top-down' mechanism where dysregulated cells outside the crypt-base stem cell niche dedifferentiate and act as tumour progenitors (Schwitalla *et al*, 2013; Davis *et al*, 2014). RNA-FISH could be used to demonstrate the spatial distribution and tissue localisation of *SHROOM2* transcripts, but may not be of similar utility for subcellular localisation. Future work should continue to focus on the generation of suitable antibodies, as it will be crucial in the investigation of *SHROOM2* function. In conjunction with this, further stable loss-of-function studies in cell lines or animal models (e.g. mice) will be beneficial, allowing more definitive phenotype characterisation before ultimately progressing to dissect the more subtle phenotypes associated with gene dosage and eQTL effects.

10.4 A hypothesis-driven approach: the genetic and non-genetic modulation of target gene (*CDH1*) expression may underlie gene-environment interactions in the predisposition to CRC

Although the eQTL analysis of colorectal mucosa and PBL has provided evidence of regulatory function for approximately half of the CRC risk loci (Chapter 6), there is still a significant proportion of risk variants whose functions and target genes are unexplained. The reasons for this could be many; in view of the gene-environment interaction (Zgaga *et al*, unpublished) between circulating vitamin D levels and the 16q22.1 risk locus, Chapter 9 outlines a hypothesis-driven approach which demonstrates that the tagging SNP rs9929218 modifies the influence of the VDR-interacting factor *FOXO4* on the target gene *CDH1*. By using expression levels derived from gene expression microarrays of the normal colorectal mucosa, multiple two-way statistical interactions were observed between rs9929218-*FOXO4* expression, *FOXO4* expression-*VDR* expression, and *FOXO4* expression-*VDR* polymorphism (*FokI*), which is in agreement with the *in silico* prediction that rs9929218 alleles possess differential FOXO-binding affinity. It implicates a biologically plausible in-vivo co-regulatory relationship between *VDR* and *FOXO4* that is modified by an established risk SNP within intron 2 of *CDH1*, providing functional support to an epidemiological gene-environment interaction.

The vitamin D active metabolite (calcitriol) treatment of CRC cell lines and human colonic organoids provided preliminary evidence that *CDH1* expression is can be induced by vitamin D activity (Chapter 9). Going forward, more detailed functional studies are required to directly demonstrate the biological interactions postulated from the observations derived from static gene expression profiles. Having now identified cell lines with a *CDH1* transcriptional response to calcitriol treatment, ligand-dependent immunoprecipitation, co-localisation, western blotting can be performed at optimal time-points to reveal post-translational modifications of FOXO4 that are mediated by VDR, as well as ligand-dependent ChIP to study DNA-protein binding at the rs9929218 locus. Additionally, gene depletion of *VDR* and *FOXO4* will establish the role of these transcription factors on *CDH1* expression, and gene reporter assays will be useful to show a ligand-dependent allele-specific differential effect of rs9929218 that mirrors the epidemiological gene-environment interaction.

To further substantiate the link between vitamin D activity, rs9929218 and *CDH1* expression, serum 25-OHD was retrospectively collected from a subset of patients who had donated colorectal tissue for gene expression profiling. There was no association with *CDH1* expression, nor were there any statistical interaction between circulating 25-OHD and rs9929218 or markers of vitamin D activity. Given that serum 25-OHD was collected at variable time points post-operatively, it may not have accurately represented the intracellular vitamin D status of the normal mucosa tissue collected during surgery. Indeed, a peri-operative time series of circulating 25-OHD examined in a prospective cohort of patients undergoing large bowel resection for CRC demonstrated a post-operative reduction of 25-OHD which did not return to pre-operative levels for at least ~5.5 months (Chapter 9). Interestingly, the time series also showed that although absolute levels of 25-OHD decreased with the surgery, relative levels were maintained at all time-points, suggesting that the inclusion of the time interval, from treatment to sampling, as a co-variate may improve statistical modelling. This finding is of importance in the wider scheme of scientific study into the effects of circulating 25-OHD and CRC outcomes, as one of the limitations of such observational studies is that the time period between surgery for the treatment of cancer and 25-OHD sampling was not constant (Zgaga *et al*, 2014).

10.5 Gender- and site-specific differential gene expression in the normal colorectal mucosa

Gene expression profiling of the normal mucosa samples used in this research revealed differences in gene expression that are influenced by gender and the anatomical site of the large bowel (Chapter 5). The gender-specific and site-specific differential expression detected in the colorectal mucosa is largely consistent with known biological processes and published literature, providing internal validation of the integrity of the samples, the microarray platform used and the processing of the data. It also demonstrates the importance of including gender and anatomical site as co-variates in the eQTL analysis to optimise the detection of subtle effects that are associated with inherited variation of gene expression.

There are known epidemiological, clinical and molecular differences between proximal and distal colon tumours, suggesting that the risk factors and transformation pathways of sporadic colorectal cancer may differ according to the anatomical location within the colon. The analysis in Chapter 5 confirms the findings of previous studies that there are widespread expression differences between the normal mucosa of the proximal and distal colorectum. Hence, it would be of interest to design future studies to analyse the proximal and distal large bowel separately (as different target tissue), in relation to heritable variation in expression (eQTL), as well as the heritable risk of CRC (GWAS). Although this will compromise the sample size, the power of the study may not necessarily suffer if there are site-specific effects that are opposing in directionality. A pilot analysis utilising the microarray data available from the samples used in this research here may inform the utility and appropriate design of such studies.

10.6 Concluding remarks

In conclusion, the work presented in this thesis has demonstrated a functional approach to discover and validate the molecular mechanisms underlying the common predisposition to colorectal cancer, and offers promise for new levels of understanding on how CRC risk variants mediates risk. The revelation that some of these common genetic variants impart risk by influencing the intermediate phenotype of transcript abundance in a tissue-specific manner adds further complexity to the study of CRC susceptibility genes and pathways. Further identification of the intermediate phenotypes for all of the risk loci will be critical in order to fully appreciate the contribution that common genetic variation makes to the development of cancer. Some of this may be achieved by examining eQTLs in specific segments of the large bowel, specific compartments/cell-types within the colonic epithelial crypt, or other tissue types altogether. Alternative intermediate phenotypes such as trans-eQTLs, long non-coding RNA, alternative transcripts and influences on the epigenome are also potential areas for future investigation. The knowledge that some of these effects may only be unveiled when analysed in relation to environmental factors highlights the need for more research into gene-environment interactions, particularly on a molecular level. Detailed understanding of the molecular consequences of inherited predisposition to this common complex disease can only have a positive impact on understanding how CRC develop and ultimately be of clinical and public health benefit.

CHAPTER 11 BIBLIOGRAPHY

Abe K, Takeichi M. EPLIN mediates linkage of the cadherin catenin complex to F-actin and stabilizes the circumferential actin belt. *Proc Natl Acad Sci U S A* 2008; 105(1): 13-9.

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. *Nat Methods* 2010; 7(4): 248-9.

Al Olama AA, Kote-Jarai Z, Giles GG, Guy M, Morrison J, Severi G, Leongamornlert DA, Tymrakiewicz M, Jhavar S, Saunders E, Hopper JL, Southey MC, Muir KR, English DR, Dearnaley DP, Ardern-Jones AT, Hall AL, O'Brien LT, Wilkinson RA, Sawyer E, Lophatananon A; UK Genetic Prostate Cancer Study Collaborators/British Association of Urological Surgeons' Section of Oncology; UK Prostate testing for cancer and Treatment study (ProtecT Study) Collaborators, Horwich A, Huddart RA, Khoo VS, Parker CC, Woodhouse CJ, Thompson A, Christmas T, Ogden C, Cooper C, Donovan JL, Hamdy FC, Neal DE, Eeles RA, Easton DF. Multiple loci on 8q24 associated with prostate cancer susceptibility. 1 *Nat Genet* 2009; 41(10): 1058-60.

Al-Tassan N, Chmiel NH, Maynard J, Fleming N, Livingston AL, Williams GT, Hodges AK, Davies DR, David SS, Sampson JR, Cheadle JP. Inherited variants of MYH associated with somatic G:C-T:A mutations in colorectal tumors. *Nat Genet* 2002; 30: 227-232.

An BS, Tavera-Mendoza LE, Dimitrov V, Wang X, Calderon MR, Wang HJ, White JH. Stimulation of Sirt1-Regulated FoxO Protein Function by the Ligand-Bound Vitamin D Receptor. *Mol Cell Biol* 2010; 30(20):4890-900

Andersen CL, Jensen JL, Orntoft TF. Normalization of real-time quantitative reverse transcription-PCR data: a model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. *Cancer Res* 2004; 64(15): 5245-50.

Anderson CL, Brown CJ. Polymorphic X-chromosome inactivation of the human TIMP1 gene. *Am J Hum Genet* 1999; 65(3): 699-708.

Anderson JW, Baird P, Davis RH, Ferreri S, Knudtson M, Koraym A, Waters V, Williams CL. Health benefits of dietary fiber. *Nutr Rev* 2009; 67: p188–205.

Arber N, Eagle CJ, Spicak J, Rácz I, Dite P, Hajer J, Zavoral M, Lechuga MJ, Gerletti P, Tang J, Rosenstein RB, Macdonald K, Bhadra P, Fowler R, Wittes J, Zauber AG, Solomon SD, Levin B; PreSAP Trial Investigators. Celecoxib for the prevention of colorectal adenomatous polyps. *N Engl J Med* 2006; 355(9): p885-95.

Arthur JC, Gharaibeh RZ, Mühlbauer M, Perez-Chanona E, Uronis JM, McCafferty J, Fodor AA, Jobin C. Microbial genomic analysis reveals the essential role of inflammation in bacteria-induced colorectal cancer. *Nat Commun* 2014; 5: 4724.

Arthur JC, Perez-Chanona E, Mühlbauer M, Tomkovich S, Uronis JM, Fan TJ, Campbell BJ, Abujamel T, Dogan B, Rogers AB, Rhodes JM, Stintzi A, Simpson KW, Hansen JJ, Keku TO, Fodor AA, Jobin C. Intestinal inflammation targets cancer-inducing activity of the microbiota. *Science* 2012; 338(6103): p120-3.

Atreya I, Neurath MF. Immune cells in colorectal cancer: prognostic relevance and therapeutic strategies. *Expert Rev Anticancer Ther* 2008; 8(4): p561-72.

Aune D, Chan DS, Lau R, Vieira R, Greenwood DC, Kampman E, Norat T. Dietary fibre, whole grains, and risk of colorectal cancer: systematic review and dose-response meta-analysis of prospective studies. *BMJ* 2011; 343: d6617.

Bai YH, Lu H, Hong D, Lin CC, Yu Z, Chen BC. Vitamin D receptor gene polymorphisms and colorectal cancer risk: a systematic metaanalysis. *World J Gastroenterol* 2012; 18(14): 1672-9.

Bailey CE, Hu CY, You YN, Bednarski BK, Rodriguez-Bigas MA, Skibber JM, Cantor SB, Chang GJ. Increasing disparities in the age-related incidences of colon and rectal cancers in the United States, 1975-2010. *JAMA Surg* 2015; 150(1): p17-22.

Barber RD, Harmer DW, Coleman RA, Clark BJ. GAPDH as a housekeeping gene: analysis of GAPDH mRNA expression in a panel of 72 human tissues. *Physiol Genomics* 2005; 21(3): 389-95.

Barker N, Ridgway RA, van Es JH, van de Wetering M, Begthel H, van den Born M, Danenberg E, Clarke AR, Sansom OJ, Clevers H. Crypt stem cells as the cells-of-origin of intestinal cancer. *Nature* 2009; 457(7229): 608-11.

Barker N, van Es JH, Kuipers J, Kujala P, van den Born M, Cozijnsen M, Haegebarth A, Korving J, Begthel H, Peters PJ, Clevers H. Identification of stem cells in small intestine and colon by marker gene *Lgr5*. *Nature* 2007; 449(7165): 1003-7.

Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 2005; 21(2): 263-5.

Bastide NM, Pierre FH, Corpet DE. Heme iron from meat and risk of colorectal cancer: a meta-analysis and a review of the mechanisms involved. *Cancer Prev Res (Phila)* 2011; 4(2): p177-84.

Batlle E, Henderson JT, Beghtel H, van den Born MM, Sancho E, Huls G, Meeldijk J, Robertson J, van de Wetering M, Pawson T, Clevers H. β-catenin and TCF mediate cell positioning in the intestinal epithelium by controlling the expression of EphB/ephrinB. *Cell* 2002;111(2): 251-63.

Beck B, Lapouge G, Rorive S, Drogat B, Desaedelaere K, Delafaille S, Dubois C, Salmon I, Willekens K, Marine JC, Blanpain C. Different levels of twist1 regulate skin tumor initiation, stemness, and progression. *Cell Stem Cell* 2015; 16(1), 67-79.

Bellam N, Pasche B. Tgf-beta signaling alterations and colon cancer. *Cancer Treat Res* 2010; 155: 85-103.

Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Statist Soc Ser 1995; B*57(1): 289-300.

Besson D, Pavageau AH, Valo I, Bourreau A, Bélanger A, Eymerit-Morin C, Moulière A, Chassevent A, Boisdron-Celle M, Morel A, Solassol J, Campone M, Gamelin E, Barré B, Coqueret O, Guette C. A quantitative proteomic approach of the different stages of colorectal cancer establishes OLFM4 as a new nonmetastatic tumor marker. *Mol Cell Proteomics* 2011; 10(12): M111.009712.

Bettington M, Walker N, Clouston A, Brown I, Leggett B, Whitehall V. The serrated pathway to colorectal carcinoma: current concepts and challenges. *Histopathology* 2013; 62(3): 367–386.

Biancolella M, Fortini BK, Tring S, Plummer SJ, Mendoza-Fandino GA, Hartiala J, Hitchler MJ, Yan C, Schumacher FR, Conti DV, Edlund CK, Noushmehr H, Coetzee SG, Bresalier RS, Ahnen DJ, Barry EL, Berman BP, Rice JC, Coetzee GA, Casey G. Identification and characterization of functional risk variants for colorectal cancer mapping to chromosome 11q23.1. *Hum Mol Genet* 2014; 23(8): 2198-209.

Birkenkamp-Demtroder K, Olesen SH, Sørensen FB, Laurberg S, Laiho P, Aaltonen LA, Orntoft TF.Differential gene expression in colon cancer of the caecum versus the sigmoid and rectosigmoid. *Gut* 2005; 54(3): 374-84.

Birkett A, Muir J, Phillips J, Jones G, O'Dea K. Resistant starch lowers fecal concentrations of ammonia and phenols in humans. *Am J Clin Nutr* 1996; 63: p766-72.

Bisgaard ML, Fenger K, Bülow S, Niebuhr E, Mohr J. Familial adenomatous polyposis (FAP): frequency, penetrance, and mutation rate. *Hum Mutat* 1994; 3(2): 121-5.

Boleij A, Hechenbleikner EM, Goodwin AC, Badani R, Stein EM, Lazarev MG, Ellis B, Carroll KC, Albesiano E, Wick EC, Platz EA, Pardoll DM, Sears CL.The Bacteroides fragilis toxin gene is prevalent in the colon mucosa of colorectal cancer patients. *Clin Infect Dis* 2015; 60(2): 208-15.

Bolognese F, Wasner M, Dohna CL, Gurtner A, Ronchi A, Muller H, Manni I, Mossner J, Piaggio G, Mantovani R, Engeland K. The cyclin B2 promoter depends on NF-Y, a trimer whose CCAAT-binding activity is cell cycle regulated. *Oncogene* 1999; 18(10): 1845-53.

Bonnet M, Buc E, Sauvanet P, Darcha C, Dubois D, Pereira B, Déchelotte P, Bonnet R, Pezet D, Darfeuille-Michaud A. Colonization of the human gut by E. coli and colorectal cancer risk. *Clin Cancer Res* 2014; 20(4): p859-67.

Botling J, Edlund K, Segersten U, Tahmasebpoor S, Engström M, Sundström M, Malmström PU, Micke P. Impact of thawing on RNA integrity and gene expression analysis in fresh frozen tissue. *Diagn Mol Pathol* 2009; 18(1): 44-52.

Botteri E, Iodice S, Bagnardi V, Raimondi S, Lowenfels AB, Maisonneuve P. Smoking and colorectal cancer: a meta-analysis. *JAMA* 2008; 300(23): p2765-78.

Boyle P and Langman JS. ABC of colorectal cancer: Epidemiology. *BMJ* 2000; 321(7264): p805-8.

Brenner H, Kloor M, Pox CP. Colorectal cancer. *Lancet* 2014; 383(9927): 1490-502.

Brenner H, Hoffmeister M, Arndt V, Haug U. Gender differences in colorectal cancer: implications for age at initiation of screening. *Br J Cancer* 2007; 96(5): 828-31.

Broderick P, Carvajal-Carmona L, Pittman AM, Webb E, Howarth K, Rowan A, Lubbe S, Spain S, Sullivan K, Fielding S, Jaeger E, Vijayakrishnan J, Kemp Z, Gorman M, Chandler

I, Papaemmanuil E, Penegar S, Wood W, Sellick G, Qureshi M, Teixeira A, Domingo E, Barclay E, Martin L, Sieber O; CORGI Consortium, Kerr D, Gray R, Peto J, Cazier JB, Tomlinson I, Houlston RS. A genome-wide association study shows that common alleles of *SMAD7* influence colorectal cancer risk. *Nat Genet* 2007; 39(11): 1315-7.

Brown CD, Mangravite LM, Engelhardt BE. Integrative Modeling of eQTLs and Cis-Regulatory Elements Suggests Mechanisms Underlying Cell Type Specificity of eQTLs. *PLoS Genet* 2013; 9(8): e1003649.

Brown CJ, Greally JM. A stain upon the silence: genes escaping X inactivation. *Trends Genet* 2003; 19(8): 432-8.

Buda A, Qualtrough D, Jepson MA, Martines D, Paraskeva C, Pignatelli M. Butyrate downregulates alpha2beta1 integrin: a possible role in the induction of apoptosis in colorectal cancer cell lines. *Gut* 2003; 52(5): p729-34.

Bullaughey K, Chavarria CI, Coop G, Gilad Y. Expression Quantitative Trait Loci detected in cell-lines are often present in primary tissues. *Hum Mol Genet* 2009; 18(22): 4296-303.

Buller RE, Sood AK, Lallas T, Buekers T, Skilling JS. Association between non-random X-chromosome inactivation and BRCA1 mutation in germline DNA of patients with ovarian cancer. *J Natl Cancer Inst* 1999; 91(4): 339-346.

Burn J, Gerdes AM, Macrae F, Mecklin JP, Moeslein G, Olschwang S, Eccles D, Evans DG, Maher ER, Bertario L, Bisgaard ML, Dunlop MG, Ho JW, Hodgson SV, Lindblom A, Lubinski J, Morrison PJ, Murday V, Ramesar R, Side L, Scott RJ, Thomas HJ, Vasen HF, Barker G, Crawford G, Elliott F, Movahedi M, Pylvanainen K, Wijnen JT, Fodde R, Lynch HT, Mathers JC, Bishop DT; CAPP2 Investigators. Long-term effect of aspirin on cancer risk in carriers of hereditary colorectal cancer: an analysis from the CAPP2 randomised controlled trial. *Lancet* 2011; 378(9809): 2081-7.

Burnett-Hartman AN, Newcomb PA, Potter JD. Infectious agents and colorectal cancer: A review of Helicobacter pylori, Streptococcus bovis, JC virus, and human papillomavirus. *Cancer Epidemiol Biomarkers Prev* 2008; 17(11): p2970-9.

Bustin SA. Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays. *J Mol Endocrinol* 2000; 25(2): 169–93.

Campione M, Steinbeisser H, Schweickert A, Deissler K, van Bebber F, Lowe LA, Nowotschin S, Viebahn C, Haffter P, Kuehn MR, Blum M.. The homeobox gene Pitx2: mediator of asymmetric left-right signaling in vertebrate heart and gut looping. Development 1999; 126 (6): 1225-34.

Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 2012; 487(7407): 330-7.

Cantero-Recasens G, Fandos C, Rubio-Moscardo F, Valverde MA, Vicente R. The asthma-associated ORMDL3 gene product regulates endoplasmic reticulum-mediated calcium signaling and cellular stress. *Hum Mol Genet* 2010; 19(1): 111-21.

Caretti G, Salsi V, Vecchi C, Imbriano C, Mantovani R. Dynamic recruitment of NF-Y and histone acetyltransferases on cell-cycle promoters. *J Biol Chem* 2003; 278(33): 30435-40.

Carew RM, Browne MB, Hickey FB, Brazil DP. Insulin receptor substrate 2 and FoxO3a signalling are involved in E-cadherin expression and transforming growth factor-β1-induced repression in kidney epithelial cells. *FEBS J* 2011; 278(18): 3370-80.

Carew RM, Browne MB, Hickey FB, Brazil DP. Insulin receptor substrate 2 and FoxO3a signalling are involved in E-cadherin expression and transforming growth factor-β1-induced repression in kidney epithelial cells. *FEBS J* 2011; 278(18): 3370-80.

Carrel L, Clemson CM, Dunn JM, Miller AP, Hunt PA, Lawrence JB, Willard HF. X inactivation analysis and DNA methylation studies of the ubiquitin activating enzyme E1 and PCTAIRE-1 genes in human and mouse. *Hum Mol Genet* 1996; 5(3): 391-401.

Carrel L, Willard HF. X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature* 2005; 434(7031): 400-4.

Carvajal-Carmona LG, Cazier JB, Jones AM, Howarth K, Broderick P, Pittman A, Dobbins S, Tenesa A, Farrington S, Prendergast J, Theodoratou E, Barnetson R, Conti D, Newcomb P, Hopper JL, Jenkins MA, Gallinger S, Duggan DJ, Campbell H, Kerr D, Casey G, Houlston R, Dunlop M, Tomlinson I.Fine-mapping of colorectal cancer susceptibility loci at 8q23.3, 16q22.1 and 19q13.11: Refinement of association signals and use of in silico analysis to suggest functional variation and unexpected candidate target genes. *Hum Mol Genet* 2011; 20(14): 2879-88.

Castellarin M, Warren RL, Freeman JD, Dreolini L, Krzywinski M, Strauss J, Barnes R, Watson P, Allen-Vercoe E, Moore RA, Holt RA. Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma. *Genome Res* 2012; 22(2): p299-306.

Ceribelli M, Dolfini D, Merico D, Gatta R, Viganò AM, Pavesi G, Mantovani R. The histone-like NF-Y is a bifunctional transcription factor. *Mol Cell Biol* 2008; 28(6): 2047-58.

Chaligné R, Heard E. X-chromosome inactivation in development and cancer. *FEBS Lett* 2014; 588(15): 2514-22.

Chan AT, Ogino S, Fuchs CS. Aspirin and the risk of colorectal cancer in relation to the expression of COX-2. *N Engl J Med* 2007; 356(21): 2131-42.

Chan DS, Lau R, Aune D, Vieira R, Greenwood DC, Kampman E, Norat T. Red and processed meat and colorectal cancer incidence: meta-analysis of prospective studies. *PLoS One* 2011; 6(6): e20456.

Chang PV, Hao L, Offermanns S, Medzhitov R. The microbial metabolite butyrate regulates intestinal macrophage function via histone deacetylase inhibition. *Proc Natl Acad Sci U S A* 2014; 111(6): p2247-52.

Cheetham SW, Gruhl F, Mattick JS, Dinger ME. Long noncoding RNAs and the genetics of cancer. *Br J Cancer* 2013; 108(12): 2419-25.

Cheung VG, Conlin LK, Weber TM, Arcaro M, Jen KY, Morley M, Spielman RS. Natural variation in human gene expression assessed in lymphoblastoid cells. *Nat Genet* 2003; 33(3): 422-5.

Cheung VG, Spielman RS. Genetics of human gene expression: mapping DNA variants that influence gene expression. *Nat Rev Genet* 2009; 10(9): 595-604.

Chew SS, Lubowski DZ. Clostridium septicum and malignancy. *ANZ J Surg* 2001; 71(11): p647-9.

Chinnappan D, Xiao D, Ratnasari A, Andry C, King TC, Weber HC. Transcription factor YY1 expression in human gastrointestinal cancer cells. *Int J Oncol* 2009; 34(5): 1417-23.

Chuaqui RF, Bonner RF, Best CJ, Gillespie JW, Flaig MJ, Hewitt SM, Phillips JL, Krizman DB, Tangrea MA, Ahram M, Linehan WM, Knezevic V, Emmert-Buck MR. Post-analysis follow-up and validation of microarray experiments. *Nat Genet* 2002; 32 Suppl: 509-14.

Chung S, Nakagawa H, Uemura M, Piao L, Ashikawa K, Hosono N, Takata R, Akamatsu S, Kawaguchi T, Morizono T, Tsunoda T, Daigo Y, Matsuda K, Kamatani N, Nakamura Y, Kubo M. Association of a novel long non-coding RNA in 8q24 with prostate cancer susceptibility. Cancer Sci 2011; 102(1): 245-52.

Church LD, Cook GP, McDermott MF. Primer: inflammasomes and interleukin 1beta in inflammatory disorders. *Nat Clin Pract Rheumatol* 2008 Jan; 4(1): 34-42.

Clarke JM, Topping DL, Bird AR, Young GP, Cobiac L. Effects of high-amylose maize starch and butyrylated high-amylose maize starch on azoxymethane-induced intestinal cancer in rats. *Carcinogenesis* 2008; 29(11): p2190-4.

Clevers H. At the crossroads of inflammation and cancer. *Cell* 2004; 118(6): p671-4.

Closa A, Cordero D, Sanz-Pamplona R, Solé X, Crous-Bou M, Paré-Brunet L, Berenguer A, Guino E, Lopez-Doriga A, Guardiola J, Biondo S, Salazar R, Moreno V. Identification of candidate susceptibility genes for colorectal cancer through eQTL analysis. *Carcinogenesis* 2014; 35(9): 2039-46.

COGENT Study, Houlston RS, Webb E, Broderick P, Pittman AM, Di Bernardo MC, Lubbe S, Chandler I, Vijayakrishnan J, Sullivan K, Penegar S; Colorectal Cancer Association Study Consortium, Carvajal-Carmona L, Howarth K, Jaeger E, Spain SL, Walther A, Barclay E, Martin L, Gorman M, Domingo E, Teixeira AS; CoRGI Consortium, Kerr D, Cazier JB, Niittymäki I, Tuupanen S, Karhu A, Aaltonen LA, Tomlinson IP, Farrington SM, Tenesa A, Prendergast JG, Barnetson RA, Cetnarskyj R, Porteous ME, Pharoah PD, Koessler T, Hampe J, Buch S, Schafmayer C, Tepel J, Schreiber S, Völzke H, Chang-Claude J, Hoffmeister M, Brenner H, Zanke BW, Montpetit A, Hudson TJ, Gallinger S, Campbell H, Dunlop MG.Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nat Genet* 2008; 40(12): 1426-35.

COGENT study. Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nat Genet* 2008; 40(12): p1426-34.

Cohen CJ, Lock WM, Mager DL. Endogenous retroviral LTRs as promoters for human genes: a critical assessment. *Gene* 2009; 448(2): 105-14.

Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M. Mapping complex disease traits with global gene expression. *Nature Rev Genet* 2009; 10(3): 184-94.

Copois V, Bibeau F, Bascoul-Mollevi C, Salvetat N, Chalbos P, Bareil C, Candeil L, Fraslon C, Conseiller E, Granci V, Mazière P, Kramar A, Ychou M, Pau B, Martineau P, Molina F, Del Rio M. RNA degradation on gene expression profiles: assessment of different methods to reliably determine RNA quality. *J Biotechnol* 2007; 127(4): 549-59.

Cowley MJ, Cotsapas CJ, Williams RB, Chan EK, Pulvers JN, Liu MY, Luo OJ, Nott DJ, Little PF. Intra- and inter-individual genetic differences in gene expression. *Mamm Genome* 2009; 20(5): 281-95.

Crosnier C, Stamataki D, Lewis J. Organizing cell renewal in the intestine: stem cells, signals and combinatorial control. *Nat Rev Genet* 2006; 7(5): 349-59.

Cross AJ, Ferrucci LM, Risch A, Graubard BI, Ward MH, Park Y, Hollenbeck AR, Schatzkin A, Sinha R. A large prospective study of meat consumption and colorectal cancer risk: an investigation of potential mechanisms underlying this association. *Cancer Res* 2010; 70: p2406-14.

Crowther-Swanepoel D, Broderick P, Di Bernardo MC, Dobbins SE, Torres M, Mansouri M, Ruiz-Ponte C, Enjuanes A, Rosenquist R, Carracedo A, Jurlander J, Campo E, Juliusson G, Montserrat E, Smedby KE, Dyer MJ, Matutes E, Dearden C, Sunter NJ, Hall AG, Mainou-Fowler T, Jackson GH, Summerfield G, Harris RJ, Pettitt AR, Allsup DJ, Bailey JR, Pratt G, Pepper C, Fegan C, Parker A, Oscier D, Allan JM, Catovsky D, Houlston RS. Common variants at 2q37.3, 8q24.21, 15q21.3 and 16q24.1 influence chronic lymphocytic leukemia risk. *Nat Genet* 2010; 42(2): 132-6.

Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y, Gräf S, Ha G, Haffari G, Bashashati A, Russell R, McKinney S; METABRIC Group, Langerød A, Green A, Provenzano E, Wishart G, Pinder S, Watson P, Markowetz F, Murphy L, Ellis I, Purushotham A, Børresen-Dale AL, Brenton JD, Tavaré S, Caldas C, Aparicio S. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* 2012; 486(7403): 346-352.

Dalerba P, Kalisky T, Sahoo D, Rajendran PS, Rothenberg ME, Leyrat AA, Sim S, Okamoto J, Johnston DM, Qian D, Zabala M, Bueno J, Neff NF, Wang J, Shelton AA, Visser B, Hisamori S, Shimono Y, van de Wetering M, Clevers H, Clarke MF, Quake SR. Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nat Biotechnol* 2011; 29(12): 1120-7.

Davis H, Irshad S, Bansal M, Rafferty H, Boitsova T, Bardella C, Jaeger E, Lewis A, Freeman-Mills L, Giner FC, Rodenas-Cuadrado P, Mallappa S, Clark S, Thomas H, Jeffery R, Poulsom R, Rodriguez-Justo M, Novelli M, Chetty R, Silver A, Sansom OJ, Greten FR, Wang LM, East JE, Tomlinson I, Leedham SJ. Aberrant epithelial GREM1 expression initiates colonic tumorigenesis from cells outside the stem cell niche. *Nat Med* 2015; 21(1): 62-70.

de Jonge HJ, Fehrmann RS, de Bont ES, Hofstra RM, Gerbens F, Kamps WA, de Vries EG, van der Zee AG, te Meerman GJ, ter Elst A. Evidence based selection of housekeeping genes. *PLoS One* 2007; 2(9): e898.

De Sousa E, Melo F, Wang X, Jansen M, Fessler E, Trinh A, de Rooij LP, de Jong JH, de Boer OJ, van Leersum R, Bijlsma MF, Rodermond H, van der Heijden M, van Noesel CJ, Tuynman JB, Dekker E, Markowetz F, Medema JP, Vermeulen L. Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions. *Nat Med* 2013; 19(5): 614-8.

Derveaux S, Vandesompele, Hellemans J. How to do successful gene expression analysis using real-time PCR. *Methods* 2010; 50(4): 227-30.

Dheda K, Huggett JF, Chang JS, Kim LU, Bustin SA, Johnson MA, Rook GA, Zumla A. The implications of using an inappropriate reference gene for real-time reverse transcription PCR data normalization. *Anal Biochem* 2005; 344(1): 141-3.

Díaz-López A, Iniesta P, Morán A, Ortega P, Fernández-Marcelo T, Sánchez-Pernaute A, Torres AJ, Benito M, De Juan C. Expression of Human MDGA1 Increases Cell Motility and Cell-Cell Adhesion and Reduces Adhesion to Extracellular Matrix Proteins in MDCK Cells. *Cancer Microenviron* 2010; 4(1): 23-32.

Dietz ML, Bernaciak TM, Vendetti F, Kielec JM, Hildebrand JD. Differential actin-dependent localization modulates the evolutionarily conserved activity of Shroom family proteins. J Biol Chem 2006; 281 (29): 20542–54.

Dimas AS, Deutsch S, Stranger BE, Montgomery SB, Borel C, Attar-Cohen H, Ingle C, Beazley C, Gutierrez Arcelus M, Sekowska M, Gagnebin M, Nisbett J, Deloukas P, Dermitzakis ET, Antonarakis SE. Common Regulatory Variation Impacts Gene Expression in a Cell Type-Dependent Manner. *Science* 2009; 325(5945): 1246-50.

Ding EL, Mehta S, Fawzi WW, Giovannucci EL. Interaction of estrogen therapy with calcium and vitamin D supplementation on colorectal cancer risk: reanalysis of Women's Health Initiative randomized trial. *Int J Cancer* 2008; 122(8): 1690-4.

Dixon AL, Liang L, Moffatt MF, Chen W, Heath S, Wong KC, Taylor J, Burnett E, Gut I, Farrall M, Lathrop GM, Abecasis GR, Cookson WO. A genome-wide association study of global gene expression. Nature Genet 2007; 39(10): 1202-07.

Dolfini D, Mantovani R.Targeting the Y/CCAAT box in cancer: YB-1 (YBX1) or NF-Y? *Cell Death Differ* 2013; 20(5): 676-85.

Dolfini D, Zambelli F, Pavesi G, Mantovani R. A perspective of promoter architecture from the CCAAT box. *Cell Cycle* 2009; 8(24): 4127-37.

Dove WF, Clipson L, Gould KA, Luongo C, Marshall DJ, Moser AR, Newton MA, Jacoby RF. Intestinal neoplasia in the *Apc*$^{\mathrm{Min}}$ mouse: independence from the microbial and natural killer (*beige* locus) status. *Cancer Res 1997;* 57(5): 812-4.

Dunlop MG, Dobbins SE, Farrington SM, Jones AM, Palles C, Whiffin N, Tenesa A, Spain S, Broderick P, Ooi LY, Domingo E, Smillie C, Henrion M, Frampton M, Martin L, Grimes G, Gorman M, Semple C, Ma YP, Barclay E, Prendergast J, Cazier JB, Olver B, Penegar S, Lubbe S, Chander I, Carvajal-Carmona LG, Ballereau S, Lloyd A, Vijayakrishnan J, Zgaga L, Rudan I, Theodoratou E; Colorectal Tumour Gene Identification (CORGI) Consortium, Starr JM, Deary I, Kirac I, Kovacević D, Aaltonen LA, Renkonen-Sinisalo L, Mecklin JP, Matsuda K, Nakamura Y, Okada Y, Gallinger S, Duggan DJ, Conti D, Newcomb P, Hopper J, Jenkins MA, Schumacher F, Casey G, Easton D, Shah M, Pharoah P, Lindblom A, Liu T; Swedish Low-Risk Colorectal Cancer Study Group, Smith CG, West H, Cheadle JP; COIN Collaborative Group, Midgley R, Kerr DJ, Campbell H, Tomlinson IP, Houlston RS. Common variation near CDKN1A, POLD3 and SHROOM2 influences colorectal cancer risk. Nat Genet 2012; 44(7): 770-6.

Dunlop MG, Farrington SM, Bubb VJ, Cunningham C, Wright M, Curtis LJ, Butt ZA, Wright E, Fleck BW, Redhead D, Mitchell R, Rainey JB, Macintyre IM, Carter DC, Wyllie AH.. Extracolonic features of familial adenomatous polyposis in patients with sporadic colorectal cancer. *Br J Cancer* 1996: 74(11), 1789-95.

Dunlop MG, Farrington SM, Carothers AD, Wyllie AH, Sharp L, Burn J, Liu B, Kinzler KW, Vogelstein B. Cancer risk associated with germline DNA mismatch repair gene mutations. *Hum Mol Genet* 1997; 6(1): 105-10.

Dunlop MG, Tenesa A, Farrington SM, Ballereau S, Brewster DH, Koessler T, Pharoah P, Schafmayer C, Hampe J, Völzke H, Chang-Claude J, Hoffmeister M, Brenner H, von Holst S, Picelli S, Lindblom A, Jenkins MA, Hopper JL, Casey G, Duggan D, Newcomb PA, Abulí A, Bessa X, Ruiz-Ponte C, Castellví-Bel S, Niittymäki I, Tuupanen S, Karhu A, Aaltonen L, Zanke B, Hudson T, Gallinger S, Barclay E, Martin L, Gorman M, Carvajal-Carmona L, Walther A, Kerr D, Lubbe S, Broderick P, Chandler I, Pittman A, Penegar S, Campbell H, Tomlinson I, Houlston RS. Cumulative impact of common genetic variants and other risk factors on colorectal cancer risk in 42,103 individuals. *Gut* 2013; 62(6): 871-81.

Easton DF, Pooley KA, Dunning AM, Pharoah PD, Thompson D, Ballinger DG, Struewing JP, Morrison J, Field H, Luben R, Wareham N, Ahmed S, Healey CS, Bowman R; SEARCH collaborators, Meyer KB, Haiman CA, Kolonel LK, Henderson BE, Le Marchand L, Brennan P, Sangrajrang S, Gaborieau V, Odefrey F, Shen CY, Wu PE, Wang HC, Eccles D, Evans DG, Peto J, Fletcher O, Johnson N, Seal S, Stratton MR, Rahman N, Chenevix-Trench G, Bojesen SE, Nordestgaard BG, Axelsson CK, Garcia-Closas M, Brinton L, Chanock S, Lissowska J, Peplonska B, Nevanlinna H, Fagerholm R, Eerola H, Kang D, Yoo KY, Noh DY, Ahn SH, Hunter DJ, Hankinson SE, Cox DG, Hall P, Wedren S, Liu J, Low YL, Bogdanova N, Schürmann P, Dörk T, Tollenaar RA, Jacobi CE, Devilee P, Klijn JG, Sigurdson AJ, Doody MM, Alexander BH, Zhang J, Cox A, Brock IW, MacPherson G, Reed MW, Couch FJ, Goode EL, Olson JE, Meijers-Heijboer H, van den Ouweland A, Uitterlinden A, Rivadeneira F, Milne RL, Ribas G, Gonzalez-Neira A, Benitez J, Hopper JL, McCredie M, Southey M, Giles GG, Schroen C, Justenhoven C, Brauch H, Hamann U, Ko YD, Spurdle AB, Beesley J, Chen X; kConFab; AOCS Management Group, Mannermaa A, Kosma VM, Kataja V, Hartikainen J, Day NE, Cox DR, Ponder BA.Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* 2007; 447(7148): 1087-93.

Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. GOrilla: A Tool For Discovery And Visualization of Enriched GO Terms in Ranked Gene Lists. *BMC Bioinformatics* 2009, 10:48.

Edwards SL, Beesley J, French JD, Dunning AM. Beyond GWASs: Illuminating the Dark Road from Association to Function. *Am J Hum Genet* 2013; 93(5): 779-97.

Eeles RA, Olama AA, Benlloch S, Saunders EJ, Leongamornlert DA, Tymrakiewicz M, Ghoussaini M, Luccarini C, Dennis J, Jugurnauth-Little S, Dadaev T, Neal DE, Hamdy FC, Donovan JL, Muir K, Giles GG, Severi G, Wiklund F, Gronberg H, Haiman CA, Schumacher F, Henderson BE, Le Marchand L, Lindstrom S, Kraft P, Hunter DJ, Gapstur S, Chanock SJ, Berndt SI, Albanes D, Andriole G, Schleutker J, Weischer M, Canzian F, Riboli E, Key TJ, Travis RC, Campa D, Ingles SA, John EM, Hayes RB, Pharoah PD, Pashayan N, Khaw KT, Stanford JL, Ostrander EA, Signorello LB, Thibodeau SN, Schaid D, Maier C, Vogel W, Kibel AS, Cybulski C, Lubinski J, Cannon-Albright L, Brenner H, Park JY, Kaneva R, Batra J, Spurdle AB, Clements JA, Teixeira MR, Dicks E, Lee A, Dunning AM, Baynes C, Conroy D, Maranian MJ, Ahmed S, Govindasami K, Guy M, Wilkinson RA, Sawyer EJ, Morgan A, Dearnaley DP, Horwich A, Huddart RA, Khoo VS, Parker CC, Van As NJ, Woodhouse CJ, Thompson A, Dudderidge T, Ogden C, Cooper CS, Lophatananon A, Cox A, Southey MC, Hopper JL, English DR, Aly M, Adolfsson J, Xu J, Zheng SL, Yeager M, Kaaks R, Diver WR, Gaudet MM, Stern MC, Corral R, Joshi AD, Shahabi A, Wahlfors T, Tammela TL, Auvinen A, Virtamo J, Klarskov P, Nordestgaard BG, Røder MA, Nielsen

SF, Bojesen SE, Siddiq A, Fitzgerald LM, Kolb S, Kwon EM, Karyadi DM, Blot WJ, Zheng W, Cai Q, McDonnell SK, Rinckleb AE, Drake B, Colditz G, Wokolorczyk D, Stephenson RA, Teerlink C, Muller H, Rothenbacher D, Sellers TA, Lin HY, Slavov C, Mitev V, Lose F, Srinivasan S, Maia S, Paulo P, Lange E, Cooney KA, Antoniou AC, Vincent D, Bacot F, Tessier DC; COGS–Cancer Research UK GWAS–ELLIPSE (part of GAME-ON) Initiative; Australian Prostate Cancer Bioresource; UK Genetic Prostate Cancer Study Collaborators/British Association of Urological Surgeons' Section of Oncology; UK ProtecT (Prostate testing for cancer and Treatment) Study Collaborators; PRACTICAL (Prostate Cancer Association Group to Investigate Cancer-Associated Alterations in the Genome) Consortium, Kote-Jarai Z, Easton DF. Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array. *Nat Genet* 2013; 45(4): 385-91, 3911-2.

Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* 1998; 95(25): 14863-8.

Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* 1998; 95(25): 14863-68.

Elkon R, Linhart C, Sharan R, Shamir R, Shiloh Y. Genome-wide in silico identification of transcriptional regulators controlling the cell cycle in human cells. *Genome Res* 2003; 13(5): 773-80.

Emilsson V. Genetics of gene expression and its effect on disease. Nature 2008; 452(7186); 423-428 .

Etournay R, Zwaenepoel I, Perfettini I, Legrain P, Petit C, El-Amraoui A. Shroom2, a myosin-VIIa- and actin-binding protein, directly interacts with ZO-1 at tight junctions. *J Cell Sci* 2007; 120(Pt 16): 2838-50.

Fairbank PD, Lee C, Ellis A, Hildebrand JD, Gross JM, Wallingford JB. Shroom2 (APXL) regulates melanosome biogenesis and localization in the retinal pigment epithelium. *Development* 2006; 133(20): 4109-18.

Farber MJ, Rizaldy R, Hildebrand JD. Shroom2 regulates contractility to control endothelial morphogenesis. *Mol Biol Cell* 2011; 22(6): 795-805.

Farber MJ. *Shroom2 regulates endothelial morphogenesis and centrosome duplication through the specific sub-cellular recruitment of Rho-kinase*. Doctoral Dissertation. University of Pittsburgh; 2012.

Farrington SM, Tenesa A, Barnetson R, Wiltshire A, Prendergast J, Porteous M, Campbell H, Dunlop MG. Germline susceptibility to colorectal cancer due to base-excision repair gene defects. *Am J Hum Genet* 2005; 77(1): 112-9.

Fearon ER, Vogelstein B. A genetic model for colorectal tumorigenesis. *Cell* 1990; 61: p759-67.

Feskanich D, Ma J, Fuchs CS, Kirkner GJ, Hankinson SE, Hollis BW, Giovannucci EL. Plasma vitamin D metabolites and risk of colorectal cancer in women. *Cancer Epidemiol Biomarkers Prev* 2004; 13(9): 1502-8.

Fitzgerald S, Sheehan KM, Espina V, O'Grady A, Cummins R, Kenny D, Liotta L, O'Kennedy R, Kay EW, Kijanka GS. High CerS5 expression levels associate with reduced patient survival and transition from apoptotic to autophagy signalling pathways in colorectal cancer. *J Path: Clin Res* 2015; 1: 54-65.

Flossmann E, Rothwell PM; British Doctors Aspirin Trial and the UK-TIA Aspirin Trial. Effect of aspirin on long-term risk of colorectal cancer: consistent evidence from randomised and observational studies. *Lancet* 2007; 369(9573): p1603-13.

Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, Jia M, Shepherd R, Leung K, Menzies A, Teague JW, Campbell PJ, Stratton MR, Futreal PA. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res* 2011; 39(Database issue): D945-50.

Foulkes WD. Inherited susceptibility to common cancers. *N Engl J Med 2008;* 359 (20): 2143-53.

Fu J, Wolfs MG, Deelen P, Westra HJ, Fehrmann RS, Te Meerman GJ, Buurman WA, Rensen SS, Groen HJ, Weersma RK, van den Berg LH, Veldink J, Ophoff RA, Snieder H, van Heel D, Jansen RC, Hofker MH, Wijmenga C, Franke L. Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. *PLoS Genet* 2012; 8(1): e1002431.

Fung KY, Ooi CC, Zucker MH, Lockett T, Williams DB, Cosgrove LJ, Topping DL. Colorectal carcinogenesis: a cellular response to sustained risk environment. *Int J Mol Sci* 2013; 14: p13525-41.

Gandini S, Boniol M, Haukka J, Byrnes G, Cox B, Sneyd MJ, Mullie P, Autier P. Meta-analysis of observational studies of serum 25-hydroxyvitamin D levels and colorectal, breast and prostate cancer and colorectal adenoma. *Int J Cancer* 2011; 128(6): 1414-24.

Ganem NJ, Godinho SA, Pellman D. A mechanism linking extra centrosomes to chromosomal instability. *Nature* 2009; 460(7252): 278-82.

Ganem NJ, Godinho SA, Pellman D. A mechanism linking extra centrosomes to chromosomal instability. *Nature* 2009; 460(7252): 278-82

Garland CF, Comstock GW, Garland FC, Helsing KJ, Shaw EK, Gorham ED. Serum 25-hydroxyvitamin D and colon cancer: eight-year prospective study. *Lancet* 1989; 2(8673): 1176-8.

Garland CF, Garland FC. Do sunlight and vitamin D reduce the likelihood of colon cancer? *Int J Epidemiol* 1980; 9(3): 227-31.

Gerbe F, van Es JH, Makrini L, Brulin B, Mellitzer G, Robine S, Romagnolo B, Shroyer NF, Bourgaux JF, Pignodel C, Clevers H, Jay P. Distinct ATOH1 and Neurog3 requirements define tuft cells as a new secretory cell type in the intestinal epithelium. *J Cell Biol 2011,* 192(5): 767-80.

Gersemann M, Becker S, Nuding S, Antoni L, Ott G, Fritz P, Oue N, Yasui W, Wehkamp J, Stange EF. Olfactomedin-4 is a glycoprotein secreted into mucus in active IBD. *J Crohns Colitis* 2012; 6(4): 425-34.

Ghoussaini M, Song H, Koessler T, Al Olama AA, Kote-Jarai Z, Driver KE, Pooley KA, Ramus SJ, Kjaer SK, Hogdall E, DiCioccio RA, Whittemore AS, Gayther SA, Giles GG, Guy M, Edwards SM, Morrison J, Donovan JL, Hamdy FC, Dearnaley DP, Ardern-Jones AT, Hall AL, O'Brien LT, Gehr-Swain BN, Wilkinson RA, Brown PM, Hopper JL, Neal DE, Pharoah PD, Ponder BA, Eeles RA, Easton DF, Dunning AM; UK Genetic Prostate Cancer Study Collaborators/British Association of Urological Surgeons' Section of Oncology; UK ProtecT Study Collaborators. Multiple loci with different cancer specificities within the 8q24 gene desert. *J Natl Cancer Inst* 2008; 100(13): 962-6.

Giovannucci E. The epidemiology of vitamin D and cancer incidence and mortality: a review (United States). *Cancer Causes Control* 2005; 16(2): 83-95.

Giovannucci E. The role of vitamin D in cancer incidence and mortality. American Association for Cancer Research Annual Meeting, Anaheim CA, 2005.

Giovannucci E. Vitamin D and cancer incidence in the Harvard cohorts. *Ann Epidemiol* 2009; 19(2): 84-8.

Glass D, Viñuela A, Davies MN, Ramasamy A, Parts L, Knowles D, Brown AA, Hedman AK, Small KS, Buil A, Grundberg E, Nica AC, Di Meglio P, Nestle FO, Ryten M; UK Brain Expression consortium; MuTHER consortium, Durbin R, McCarthy MI, Deloukas P, Dermitzakis ET, Weale ME, Bataille V, Spector TD. Gene expression changes with age in skin, adipose tissue, blood and brain. *Genome Biol* 2013; 14(7): R75.

Glebov OK, Rodriguez LM, Nakahara K, Jenkins J, Cliatt J, Humbyrd CJ, DeNobile J, Soballe P, Simon R, Wright G, Lynch P, Patterson S, Lynch H, Gallinger S, Buchbinder A, Gordon G, Hawk E, Kirsch IR.Distinguishing Right from Left Colon by the Pattern of Gene Expression. *Cancer Epidemiol Biomarkers Prev* 2003; 12(8): 755-62.

Gordon D, Abajian C, Green P. Consed: A Graphical Tool for Sequence Finishing. Genome Res 1998; 8(3): 195-202.

Gorham ED, Garland CF, Garland FC, Grant WB, Mohr SB, Lipkin M, Newmark HL, Giovannucci E, Wei M, Holick MF. Vitamin D and prevention of colorectal cancer.

Goring HH, Curran JE, Johnson MP, Dyer TD, Charlesworth J, Cole SA, Jowett JB, Abraham LJ, Rainwater DL, Comuzzie AG, Mahaney MC, Almasy L, MacCluer JW, Kissebah AH, Collier GR, Moses EK, Blangero J. Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nature Genet* 2007; 39(10): 1208-16.

Goss KH, Groden J. Biology of the adenomatous polyposis coli tumor suppressor. *J Clin Oncol* 2000; 18(9):1967-79.

Graham RR, Kozyrev SV, Baechler EC, Reddy MV, Plenge RM, Bauer JW, Ortmann WA, Koeuth T, González Escribano MF, Argentine and Spanish Collaborative Groups, Pons-Estel B, Petri M, Daly M, Gregersen PK, Martín J, Altshuler D, Behrens TW, Alarcón-Riquelme ME. A common haplotype of interferon regulatory factor 5 (IRF5) regulates splicing and expression and is associated with increased risk of systemic lupus erythematosus. *Nat Genet* 2006; 38(5): 550-5.

Grant WB, Garland CF. A critical review of studies on vitamin D in relation to colorectal cancer. *Nutr Cancer* 2004; 48(2): 115-23.

Grundberg, Small KS, Hedman ÅK, Nica AC, Buil A, Keildson S, Bell JT, Yang TP, Meduri E, Barrett A, Nisbett J, Sekowska M, Wilk A, Shin SY, Glass D, Travers M, Min JL, Ring S, Ho K, Thorleifsson G, Kong A, Thorsteindottir U, Ainali C, Dimas AS, Hassanali N, Ingle C, Knowles D, Krestyaninova M, Lowe CE, Di Meglio P, Montgomery SB, Parts L, Potter S, Surdulescu G, Tsaprouni L, Tsoka S, Bataille V, Durbin R, Nestle FO, O'Rahilly S, Soranzo N, Lindgren CM, Zondervan KT, Ahmadi KR, Schadt EE, Stefansson K, Smith GD, McCarthy MI, Deloukas P, Dermitzakis ET, Spector TD; Multiple Tissue Human Expression Resource (MuTHER) Consortium. Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nature Genet* 2012; 44(10): 1084-9.

Guelen L, Pagie L, Brasset E, Meuleman W, Faza MB, Talhout W, Eussen BH, de Klein A, Wessels L, de Laat W, van Steensel B. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* 2008; 453(7197): 948-951.

Gutschner T, Diederichs S. The hallmarks of cancer: a long non-coding RNA point of view. *RNA Biol* 2012; 9(6): 703-19.

Ha SG, Ge XN, Bahaie NS, Kang BN, Rao A, Rao SP, Sriramarao P. ORMDL3 promotes eosinophil trafficking and activation via regulation of integrins and CD48. *Nat Commun* 2013; 4: 2479.

Halbrooks PJ, Ding R, Wozney JM, Bain G. Role of RGM coreceptors in bone morphogenetic protein signaling. *J Mol Signal* 2007; 2: 4.

Halbrooks PJ, Ding R, Wozney JM, Bain G. Role of RGM coreceptors in bone morphogenetic protein signaling. *J Mol Signal* 2007; 2: 4.

Half E, Arber N. Colon cancer: preventive agents and the present status of chemoprevention. *Expert Opin Pharmacother* 2009 Feb; 10(2): p211-9.

Hall A. The cytoskeleton and cancer. *Cancer Metastasis Rev* 2009; 28(1-2): 5-14.

Hardwick JC, Kodach LL, Offerhaus GJ, van den Brink GR. Bone morphogenetic protein signalling in colorectal cancer. *Nat Rev Cancer* 2008; 8(10): 806-12.

Harismendy O, Notani D, Song X, Rahim NG, Tanasa B, Heintzman N, Ren B, Fu XD, Topol EJ, Rosenfeld MG, Frazer KA. 9p21 DNA variants associated with coronary artery disease impair interferon-γ signalling response. *Nature* 2011; 470(7333): 264-8.

Harper LV, Hilton AC, Jones AF. RT-PCR for the pseudogene-free amplification of the glyceraldehyde-3-phosphate dehydrogenase gene (gapd). *Mol Cell Probes* 2003; 17(5): 261-5.

Hatakeyama C, Anderson CL, Beever CL, Peñaherrera MS, Brown CJ, Robinson WP. The dynamics of X-inactivation skewing as women age. *Clin Genet* 2004; 66(4): 327-32.

Hawthorn L, Lan L, Mojica W. Evidence for field effect cancerization in colorectal cancer. *Genomics* 2014; 103(2-3): 211-21.

He TC, Sparks AB, Rago C, Hermeking H, Zawel L, da Costa LT, Morin PJ, Vogelstein B, Kinzler KW. Identification of c-MYC as a target of the APC pathway. *Science* 1998; 281(5382): 1509-12.

He XC, Zhang J, Tong WG, Tawfik O, Ross J, Scoville DH, Tian Q, Zeng X, He X, Wiedemann LM, Mishina Y, Li L. BMP signaling inhibits intestinal stem cell self-renewal through suppression of Wnt–β-catenin signaling. *Nature Genet* 2005; 36(10): 1117-21.

Heinen CD. Genotype to Phenotype: Analyzing the Effects of Inherited Mutations in Colorectal Cancer Families. *Mutat Res* 2010; 693(1-2): 32-45.

Herman JG, Umar A, Polyak K, Graff JR, Ahuja N, Issa JP, Markowitz S, Willson JK, Hamilton SR, Kinzler KW, Kane MF, Kolodner RD, Vogelstein B, Kunkel TA, Baylin SB. Incidence and functional consequences of hMLH1 promoter hypermethylation in colorectal carcinoma. *Proc Natl Acad Sci USA* 1998; 95(12): 6870-75.

Hinoue T, Weisenberger DJ, Lange CP, Shen H, Byun HM, Van Den Berg D, Malik S, Pan F, Noushmehr H, van Dijk CM, Tollenaar RA, Laird PW. Genome-scale analysis of aberrant DNA methylation in colorectal cancer. *Genome Res* 2012; 22(2): 271-82.

Hirose H, Ishii H, Mimori K, Tanaka F, Takemasa I, Mizushima T, Ikeda M, Yamamoto H, Sekimoto M, Doki Y, Mori M. The significance of PITX2 overexpression in human colorectal cancer. *Ann Surg Oncol* 2011; 18(10): 3005-12.

Holick MF. Photobiology of vitamin D. In: Feldman D, Glorieux FH, Pike JW (eds.) Vitamin D. San Diego: Academic Press; 1997. p33-9.

Houlston RS, Cheadle J, Dobbins SE, Tenesa A, Jones AM, Howarth K, Spain SL, Broderick P, Domingo E, Farrington S, Prendergast JG, Pittman AM, Theodoratou E, Smith CG, Olver B, Walther A, Barnetson RA, Churchman M, Jaeger EE, Penegar S, Barclay E, Martin L, Gorman M, Mager R, Johnstone E, Midgley R, Niittymäki I, Tuupanen S, Colley J, Idziaszczyk S; COGENT Consortium, Thomas HJ, Lucassen AM, Evans DG, Maher ER; CORGI Consortium; COIN Collaborative Group; COINB Collaborative Group, Maughan T, Dimas A, Dermitzakis E, Cazier JB, Aaltonen LA, Pharoah P, Kerr DJ, Carvajal-Carmona LG, Campbell H, Dunlop MG, Tomlinson IP. Meta-analysis of three genome-wide association studies identifies susceptibility loci for colorectal cancer at 1q41, 3q26.2, 12q13.13 and 20q13.33. *Nat Genet* 2010; 42(11): p973-7.

Houlston RS, Fallon T, Harocopos C, Williams CB, Davey C, Slack J. Congenital hypertrophy of retinal pigment epithelium in patients with colonic polyps associated with cancer family syndrome. *Clin Genet* 1992; 42(1): 16-8.

Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nature Protoc* 2009; 4(1):44-57.

Huang J, Qi R, Quackenbush J, Dauway E, Lazaridis E, Yeatman T. Effects of ischemia on gene expression. *J Surg Res* 2001; 99(2): 222-7.

Huggett J, Dheda K, Bustin S, Zumla A. Real-time RT-PCR normalisation; strategies and considerations. *Genes Immun* 2005; 6(4): 279-84.

Hughes JD, Estep PW, Tavazoie S, Church GM. Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J Mol Biol* 2000; 296(5): 1205-14.

Hughes R, Rowland IR. Metabolic Activities of the Gut Microflora in Relation to Cancer. *Microb Ecol Health Dis 2000*; 12(2): p179-85.

Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD, Kidd MJ, King AM, Meyer MR, Slade D, Lum PY, Stepaniants SB, Shoemaker DD, Gachotte D, Chakraburtty K, Simon J, Bard M, Friend SH. Functional discovery via a compendium of expression profiles. *Cell* 2000; 102(1):109-126.

Iacopetta B. Are there two sides to colorectal cancer? *Int J Cancer* 2002; 101(5): 403-8.
J Steroid Biochem Mol Biol 2005; 97(1-2): 179-94.

Jacob F, Guertler R, Naim S, Nixdorf S, Fedier A, Hacker NF, Heinzelmann-Schwarz V. Careful Selection of Reference Genes Is Required for Reliable Performance of RT-qPCR in Human Normal and Cancer Cell Lines. *PLoS ONE* 2013; 8:e59180.

Jacobs ET, Martínez ME, Jurutka PW. Vitamin D: Marker or Mechanism of Action? *Cancer Epidemiol Biomarkers Prev* 2011; 20(4): 585–590.

Jaeger E, Webb E, Howarth K, Carvajal-Carmona L, Rowan A, Broderick P, Walther A, Spain S, Pittman A, Kemp Z, Sullivan K, Heinimann K, Lubbe S, Domingo E, Barclay E, Martin L, Gorman M, Chandler I, Vijayakrishnan J, Wood W, Papaemmanuil E, Penegar S, Qureshi M; CORGI Consortium, Farrington S, Tenesa A, Cazier JB, Kerr D, Gray R, Peto J, Dunlop M, Campbell H, Thomas H, Houlston R, Tomlinson I. Common genetic variants at the *CRAC1* (*HMPS*) locus on chromosome 15q13.3 influence colorectal cancer risk. *Nat Genet* 2008; 40(1): 26-8.

Järvinen HJ, Aarnio M, Mustonen H, Aktan-Collan K, Aaltonen LA, Peltomäki P, De La Chapelle A, Mecklin JP. Controlled 15-year trial on screening for colorectal cancer in families with hereditary nonpolyposis colorectal cancer. *Gastroenterology* 2000; 118: p829–34.

Jendrzejewski J, He H, Radomska HS, Li W, Tomsic J, Liyanarachchi S, Davuluri RV, Nagy R, de la Chapelle A. The polymorphism rs944289 predisposes to papillary thyroid carcinoma through a large intergenic noncoding RNA gene of tumor suppressor type. *Proc Natl Acad Sci USA* 2012; 109(22): 8646–8651

Jensen J, Pedersen EE, Galante P, Hald J, Heller RS, Ishibashi M, Kageyama R, Guillemot F, Serup P, Madsen OD. Control of endodermal endocrine development by Hes-1. *Nat Genet* 2000; 24(1): 36-44.

Jiang WG, Martin TA, Lewis-Russell JM, Douglas-Jones A, Ye L, Mansel RE. Eplin-alpha expression in human breast cancer, the impact on cellular migration and clinical outcome. *Mol Cancer* 2008;7:71.

Johns LE, Houlston RS. A systematic review and meta-analysis of familial colorectal cancer risk. *Am J Gastroenterol* 2001; 96(10): 2992-3003.

Jones S, Emmerson P, Maynard J, Best JM, Jordan S, Williams GT, Sampson JR, Cheadle JP. Biallelic germline mutations in MYH predispose to multiple colorectal adenoma and somatic G:C-->T:A mutations. *Hum Mol Genet* 2002; 11(23): 2961-7.

Jordan A Roberts, Lindsay Waters, Jae Y Ro, Qihui Jim Zhai. Smoothelin and caldesmon are reliable markers for distinguishing muscularis propria from desmoplasia: a critical distinction for accurate staging colorectal adenocarcinoma. *Int J Clin Exp Pathol* 2014; 7(2): 792-6.

Jürchott K, Kuban RJ, Krech T, Blüthgen N, Stein U, Walther W, Friese C, Kiełbasa SM, Ungethüm U, Lund P, Knösel T, Kemmner W, Morkel M, Fritzmann J, Schlag PM, Birchmeier W, Krueger T, Sperling S, Sers C, Royer HD, Herzel H, Schäfer R. Identification of Y-Box Binding Protein 1 As a Core Regulator of MEK/ERK Pathway-Dependent Gene Signatures in Colorectal Cancer Cells. *PLoS Genet* 2010; 6(12): e1001231.

Kathiresan S, Melander O, Guiducci C, Surti A, Burtt NP, Rieder MJ, Cooper GM, Roos C, Voight BF, Havulinna AS, Wahlstrand B, Hedner T, Corella D, Tai ES, Ordovas JM, Berglund G, Vartiainen E, Jousilahti P, Hedblad B, Taskinen MR, Newton-Cheh C, Salomaa V, Peltonen L, Groop L, Altshuler DM, Orho-Melander M. Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nat Genet* 2008; 40(2): 189-97.

Kawasaki Y, Sato R, Akiyama T. Mutated APC and Asef are involved in the migration of colorectal tumour cells. *Nat Cell Biol* 2003; 5(3): 211-15.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. *Genome Res* 2002;12(6): 996-1006.

Kiemeney LA, Thorlacius S, Sulem P, Geller F, Aben KK, Stacey SN, Gudmundsson J, Jakobsdottir M, Bergthorsson JT, Sigurdsson A, Blondal T, Witjes JA, Vermeulen SH, Hulsbergen-van de Kaa CA, Swinkels DW, Ploeg M, Cornel EB, Vergunst H, Thorgeirsson TE, Gudbjartsson D, Gudjonsson SA, Thorleifsson G, Kristinsson KT, Mouy M, Snorradottir S, Placidi D, Campagna M, Arici C, Koppova K, Gurzau E, Rudnai P, Kellen E, Polidoro S, Guarrera S, Sacerdote C, Sanchez M, Saez B, Valdivia G, Ryk C, de Verdier P, Lindblom A, Golka K, Bishop DT, Knowles MA, Nikulasson S, Petursdottir V, Jonsson E, Geirsson G, Kristjansson B, Mayordomo JI, Steineck G, Porru S, Buntinx F, Zeegers MP, Fletcher T, Kumar R, Matullo G, Vineis P, Kiltie AE, Gulcher JR, Thorsteinsdottir U, Kong A, Rafnar T, Stefansson K. Sequence variant on 8q24 confers susceptibility to urinary bladder cancer. *Nat Genet* 2008; 40(11): 1307-12.

Kim SK, Lund J, Kiraly M, Duke K, Jiang M, Stuart JM, Eizinger A, Wylie BN, Davidson GS. A gene expression map for Caenorhabditis elegans. *Science* 2001; 293(5537): 2087-92.

Kim SK, Lund J, Kiraly M, Duke K, Jiang M, Stuart JM, Eizinger A, Wylie BN, Davidson GS. A gene expression map for Caenorhabditis elegans. *Science* 2001; 293(5537): 2087-2092.

Kinzler KW, Vogelstein B. Landscaping the cancer terrain. *Science* 1998; 280(5366): 1036-7.

Kioussi C, Briata P, Baek SH, Rose DW, Hamblet NS, Herman T, Ohgi KA, Lin C, Gleiberman A, Wang J, Brault V, Ruiz-Lozano P, Nguyen HD, Kemler R, Glass CK, Wynshaw-Boris A, Rosenfeld MG. Identification of a Wnt/Dvl/beta-Catenin --> Pitx2 pathway mediating cell-type-specific proliferation during development. Cell 2002; 111(5): 673-85.

Knox S, Harris J, Calton L, Wallace AM. A simple automated solid-phase extraction procedure for measurement of 25-hydroxyvitamin D3 and D2 by liquid chromatography-tandem mass spectrometry. *Ann Clin Biochem* 2009; 46(Pt 3): 226-30.

Kolodner RD, Tytell JD, Schmeits JL, Kane MF, Gupta RD, Weger J, Wahlberg S, Fox EA, Peel D, Ziogas A, Garber JE, Syngal S, Anton-Culver H, Li FP. Germ-line msh6 mutations in colorectal cancer families. *Cancer Res* 1999; 59(20): 5068-74.

Koo JH, Leong RWL. Sex Differences in Epidemiological, Clinical and Pathological Characteristics of Colorectal Cancer. *J Gastroenterol Hepatol* 2010; 25(1): 33-42.

Korinek V, Barker N, Moerer P, van Donselaar E, Huls G, Peters PJ, Clevers H.Depletion of epithelial stem-cell compartments in the small intestine of mice lacking Tcf-4. *Nat Genet 1998;* 19(4): 379-83.

Kosinski C, Li VS, Chan AS, Zhang J, Ho C, Tsui WY, Chan TL, Mifflin RC, Powell DW, Yuen ST, Leung SY, Chen X. Gene expression patterns of human colon tops and basal crypts and BMP antagonists as intestinal stem cell niche factors. *Proc Natl Acad Sci USA 2007;* 104(39): 15418-23.

Kostic AD, Gevers D, Pedamallu CS, Michaud M, Duke F, Earl AM, Ojesina AI, Jung J, Bass AJ, Tabernero J, Baselga J, Liu C, Shivdasani RA, Ogino S, Birren BW, Huttenhower C, Garrett WS, Meyerson M.Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. *Genome Res* 2012; 22(2): p292-8.

Krishnan A, Ochola J, Mundy J, Jones M, Kruger P, Duncan E, Venkatesh B. Acute fluid shifts influence the assessment of serum vitamin D status in critically ill patients. *Crit Care* 2010; 14(6): R216.

Kristiansen M, Knudsen GP, Maguire P, Margolin S, Pedersen J, Lindblom A, Ørstavik KH: High incidence of skewed X chromosome inactivation in young patients with familial non-BRCA1/BRCA2 breast cancer. *J Med Genet* 2005; 42(11): 877-880.

Kristiansen M, Langerod A, Knudsen GP, Weber BL, Borresen-Dale A-L, Ørstavik KH: High frequency of skewed X inactivation in young breast cancer patients. *J Med Genet* 2002, 39(1):30-33.

Kubista M, Andrade JM, Bengtsson M, Forootan A, Jonak J, Lind K, Sindelka R, Sjöback R, Sjögreen B, Strömbom L, Ståhlberg A, Zoric N. The real-time polymerase chain reaction. *Mol Aspects Med* 2006; 27(2-3): 95-125.

Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* 2009; 4(7): 1073-81.

Kumar V, Westra HJ, Karjalainen J, Zhernakova DV, Esko T, Hrdlickova B, Almeida R, Zhernakova A, Reinmaa E, Võsa U, Hofker MH, Fehrmann RS, Fu J, Withoff S, Metspalu A, Franke L, Wijmenga C. Human disease-associated genetic variation impacts large intergenic non-coding RNA expression. *PLoS Genet* 2013; 9(1): e1003201.

Lakatos PL, Lakatos L. Risk for colorectal cancer in ulcerative colitis: Changes, causes and management strategies. *World J Gastroenterol* 2008; 14(25): p3937-47.

Laken SJ, Petersen GM, Gruber SB, Oddoux C, Ostrer H, Giardiello FM, Hamilton SR, Hampel H, Markowitz A, Klimstra D, Jhanwar S, Winawer S, Offit K, Luce MC, Kinzler KW, Vogelstein B. Familial colorectal cancer in Ashkenazim due to a hypermutable tract in APC. *Nat Genet* 1997; *17(1):* 79-83.

Lalonde E, Ha KC, Wang Z, Bemmo A, Kleinman CL, Kwan T, Pastinen T, Majewski J. RNA sequencing reveals the role of splicing polymorphisms in regulating human gene expression. *Genome Res* 2011; 21(4): 545-54.

LaPointe LC, Dunne R, Brown GS, Worthley DL, Molloy PL, Wattchow D, Young GP.Map of differential transcript expression in the normal human large intestine. *Physiol Genomics* 2008; 33(1): 50-64.

Larsson SC, Wolk A. Meat consumption and risk of colorectal cancer: A meta-analysis of prospective studies. *Int J Cancer* 2006; 119(11): 2657-64.

Larsson SC, Giovannucci E, Wolk A. Long-term aspirin use and colorectal cancer risk: a cohort study in Sweden. *Br J Cancer* 2006; 95(9): 1277-9.

Lee C, Le MP, Wallingford JB.The Shroom family proteins play broad roles in the morphogenesis of thickened epithelial sheets. *Dev Dyn* 2009; 238(6): 1480-91.

Lefèvre N, Corazza F, Duchateau J, Desir J, Casimir G. Sex differences in inflammatory cytokines and CD99 expression following in vitro lipopolysaccharide stimulation. Shock 2012; 38(1): 37-42.

Lengauer C, Kinzler KW, Vogelstein B. Genetic instability in colorectal cancers. *Nature* 1997; 386(6625): 623-7.

Leonel AJ, Alvarez-Leite JI. Butyrate: implications for intestinal function. *Curr Opin Clin Nutr Metab Care* 2012; 15: 474-9.

Li G, Su Q, Liu GQ, Gong L, Zhang W, Wang SF, Zhu SJ, Zhang HL, Feng YM, Zhang YH. Skewed X-chromosome inactivation of bood cells is associated with early development of lung cancer in females. *Oncol Rep* 2006, 16(4): 859-64.

Liberati C, Di Silvio A, Ottolenghi S, Mantovani R. NF-Y binding to twin CCAAT boxes: role of Q-rich domains and histone fold helices. *J Mol Biol* 1999; 285(4): 1441-55.

Libert C, Dejager L, Pinheiro I. The X chromosome in immune functions: when a chromosome makes the difference. *Nat Rev Immunol* 2010; 10(8): 594-604.

Libioulle C, Louis E, Hansoul S, Sandor C, Farnir F, Franchimont D, Vermeire S, Dewit O, de Vos M, Dixon A, Demarche B, Gut I, Heath S, Foglio M, Liang L, Laukens D, Mni M, Zelenika D, Van Gossum A, Rutgeerts P, Belaiche J, Lathrop M, Georges M. Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. *PLoS Genet* 2007; 3(4): e58.

Lin DP, Wang Y, Scherer SJ, Clark AB, Yang K, Avdievich E, Jin B, Werling U, Parris T, Kurihara N, Umar A, Kucherlapati R, Lipkin M, Kunkel TA, Edelmann W. An Msh2 Point Mutation Uncouples DNA Mismatch Repair and Apoptosis. *Cancer Res* 2004; 64(2): 517-22.

Liu W, Yan M, Liu Y, Wang R, Li C, Deng C, Singh A, Coleman WG Jr, Rodgers GP. Olfactomedin 4 down-regulates innate immunity against Helicobacter pylori infection. *Proc Natl Acad Sci U S A* 2010; 107(24): 11056-61.

Liu XF, Olsson P, Wolfgang CD, Bera TK, Duray P, Lee B, Pastan I. PRAC: a novel small nuclear protein that is specifically expressed in human prostate and colon. *Prostate 2001;* 47(2): 125–131

Logan M, Pagán-Westphal SM, Smith DM, Paganessi L, Tabin CJ. The transcription factor Pitx2 mediates situs-specific morphogenesis in response to left-right asymmetric signals. Cell 1998; 94(3): 307-17.

Loo LW, Cheng I, Tiirikainen M, Lum-Jones A, Seifried A, Dunklee LM, Church JM, Gryfe R, Weisenberger DJ, Haile RW, Gallinger S, Duggan DJ, Thibodeau SN, Casey G, Le Marchand L. cis-Expression QTL analysis of established colorectal cancer risk variants in colon tumors and adjacent normal tissue. *PLoS One* 2012; 7(2): e30477.

Louis P, Hold GL, Flint HJ. The gut microbiota, bacterial metabolites and colorectal cancer. *Nat Rev Microbiol* 2014; 12(10): p661-72.

Lubbe SJ, Di Bernardo MC, Broderick P, Chandler I, Houlston RS. Comprehensive evaluation of the impact of 14 genetic variants on colorectal cancer phenotype and risk. *Am J Epidemiol* 2012; 175(1): 1-10.

Lynch HT, Lynch J. Lynch syndrome: genetics, natural history, genetic counseling, and prevention. *J Clin Oncol* 2000; 18(21 Suppl): 19S-31S.

Lynch HT, Lynch JF. Genetics of colonic cancer. *Digestion* 1998; 59(5): 481-92.

Lynch HT, Lynch PM, Lanspa SJ, Snyder CL, Lynch JF, Boland CR. Review of the Lynch syndrome: history, molecular genetics, screening, differential diagnosis, and medicolegal ramifications. *Clin Genet* 2009; 76(1): p1-18.

Madison BB, Braunstein K, Kuizon E, Portman K, Qiao XT, Gumucio DL.Epithelial hedgehog signals pattern the intestinal crypt–villus axis. *Development* 2005; 132(2), 279-89.

Manni I, Mazzaro G, Gurtner A, Mantovani R, Haugwitz U, Krause K, Engeland K, Sacchi A, Soddu S, Piaggio G. NF-Y mediates the transcriptional inhibition of the cyclin B1, cyclin B2, and cdc25C promoters upon induced G2 arrest. *J Biol Chem* 2001; 276(8): 5570-6.

Martin Sauvageau, Loyal A Goff, Simona Lodato, Boyan Bonev, Abigail F Groff, Chiara Gerhardinger, Diana B Sanchez-Gomez, Ezgi Hacisuleyman, Eric Li, Matthew Spence, Stephen C Liapis, William Mallard, Michael Morse, Mavis R Swerdel, Michael F D'Ecclessis, Jennifer C Moore, Venus Lai, Guochun Gong, George D Yancopoulos, David Frendewey, Manolis Kellis, Ronald P Hart, David M Valenzuela, Paola Arlotta, John L Rinn. Multiple knockout mouse models reveal lincRNAs are required for life and brain development. *eLife* 2013; 2: e01749.

Matano M, Date S, Shimokawa M, Takano A, Fujii M, Ohta Y, Watanabe T, Kanai T, Sato T.Modeling colorectal cancer using CRISPR-Cas9-mediated engineering of human intestinal organoids. *Nat Med* 2015; 21(3): 256-62.

Mathelier A, Zhao X, Zhang AW, Parcy F, Worsley-Hunt R, Arenillas DJ, Buchman S, Chen CY, Chou A, Ienasescu H, Lim J, Shyr C, Tan G, Zhou M, Lenhard B, Sandelin A, Wasserman WW. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res* 2014; 42(Database issue): D142-7.

Maul R, Chang D. EPLIN, Epithelial protein lost in neoplasm. *Oncogene* 1999; 18(54): 7838-41.

Maunakea AK, Nagarajan RP, Bilenky M, Ballinger TJ, D'Souza C, Fouse SD, Johnson BE, Hong C, Nielsen C, Zhao Y, Turecki G, Delaney A, Varhol R, Thiessen N, Shchors K, Heine VM, Rowitch DH, Xing X, Fiore C, Schillebeeckx M, Jones SJ, Haussler D, Marra MA, Hirst M, Wang T, Costello JF. Conserved Role of Intragenic DNA Methylation in Regulating Alternative Promoters. Nature 2010; 466(7303): 253-7.

McDonald MJ, Wang WC, Huang HD, Leu JY. Clusters of nucleotide substitutions and insertion/deletion mutations are associated with repeat sequences. *PLoS Biol* 2001; 9: e1000622.

Meda SA, Narayanan B, Liu J, Perrone-Bizzozero NI, Stevens MC, Calhoun VD, Glahn DC, Shen L, Risacher SL, Saykin AJ, Pearlson GD. A large scale multivariate parallel ICA method reveals novel imaging-genetic relationships for Alzheimer's disease in the ADNI cohort. *Neuroimage* 2012; 60(3): 1608-21.

Menon S, Tselepis C, Anderson M. Oesophagus: Is there a gender specific response of oesophageal mucosa to acid reflux? *Gut* 2011; 60: A176

Mercer TR, Dinger ME, Mattick JS. Long noncoding RNAs: insights into function. *Nat Rev Genet* 2009; 10(3): 155-59.

Mercer TR, Gerhardt DJ, Dinger ME, Crawford J, Trapnell C, Jeddeloh JA, Mattick JS, Rinn JL. Targeted RNA sequencing reveals the deep complexity of the human transcriptome. *Nat Biotechnol* 2011; 30(1): 99-104.

Merlos-Suárez A, Barriga FM, Jung P, Iglesias M, Céspedes MV, Rossell D, Sevillano M, Hernando-Momblona X, da Silva-Diz V, Muñoz P, Clevers H, Sancho E, Mangues R, Batlle E. The intestinal stem cell signature identifies colorectal cancer stem cells and predicts disease relapse. *Cell Stem Cell* 2011; 8(5): 511-24.

Meyer LA, Broaddus RR, Lu KH. Endometrial cancer and Lynch syndrome: clinical and pathologic considerations. *Cancer Control* 2009; 16(1): 14-22.

Meyskens FL Jr, McLaren CE, Pelot D, Fujikawa-Brooks S, Carpenter PM, Hawk E, Kelloff G, Lawson MJ, Kidao J, McCracken J, Albers CG, Ahnen DJ, Turgeon DK, Goldschmid S, Lance P, Hagedorn CH, Gillen DL, Gerner EW. Difluoromethylornithine plus sulindac for the prevention of sporadic colorectal adenomas: a randomized placebo-controlled, double-blind trial. *Cancer Prev Res (Phila)* 2008; 1(1): 32-8.

Miller M, Rosenthal P, Beppu A, Mueller JL, Hoffman HM, Tam AB, Doherty TA, McGeough MD, Pena CA, Suzukawa M, Niwa M, Broide DH. ORMDL3 transgenic mice have increased airway remodeling and airway responsiveness characteristic of asthma. *J Immunol* 2014; 192(8): 3475-87.

Mitchell RJ, Farrington SM, Dunlop MG, Campbell H. Mismatch repair genes hMLH1 and hMSH2 and colorectal cancer: a HuGE review. *Am J Epidemiol* 2002; 156(10): 885-902.

Mizoue T. Ecological study of solar radiation and cancer mortality in Japan. *Health Phys* 2004; 87(5): 532-8.

Moffatt MF, Kabesch M, Liang L, Dixon AL, Strachan D, Heath S, Depner M, von Berg A, Bufe A, Rietschel E, Heinzmann A, Simma B, Frischer T, Willis-Owen SA, Wong KC, Illig T, Vogelberg C, Weiland SK, von Mutius E, Abecasis GR, Farrall M, Gut IG, Lathrop GM, Cookson WO. Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature* 2007; 448(7152); 470-73.

Mooney SM, Rajagopalan K, Williams BH, Zeng Y, Christudass CS, Li Y, Yin B, Kulkarni P, Getzenberg RH. Creatine kinase brain overexpression protects colorectal cells from various metabolic and non-metabolic stresses. *J Cell Biochem* 2011; 112(4): 1066-75.

Mori Y, Olaru AV, Cheng Y, Agarwal R, Yang J, Luvsanjav D, Yu W, Selaru FM, Hutfless S, Lazarev M, Kwon JH, Brant SR, Marohn MR, Hutcheon DF, Duncan MD, Goel A, Meltzer SJ. Novel candidate colorectal cancer biomarkers identified by methylation microarray-based scanning. *Endocr Relat Cancer* 2011; 18(4): 465-78.

Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, Cheung VG. Genetic analysis of genome-wide variation in human gene expression. *Nature* 2004; 430(7001): 743-7.

Morris EJ, Ji JY, Yang F, Di Stefano L, Herr A, Moon NS, Kwon EJ, Haigis KM, Näär AM, Dyson NJ. E2F1 represses β-catenin transcription and is antagonized by both pRB and CDK8. *Nature* 2008; 455(7212): 552-6.

Mort RL, Ford MJ, Sakaue-Sawano A, Lindstrom NO, Casadio A, Douglas AT, Keighren MA, Hohenstein P, Miyawaki A, Jackson IJ. Fucci2a: A bicistronic cell cycle reporter that allows Cre mediated tissue specific expression in mice. *Cell Cycle* 2014; 13(17): 2681-96.

Moskal A, Norat T, Ferrari P, Riboli E. Alcohol intake and colorectal cancer risk: a dose-response meta-analysis of published cohort studies. *Int J Cancer* 2007; 120(3): p664-71.

Motiwala T, Kutay H, Ghoshal K, Bai S, Seimiya H, Tsuruo T, Suster S, Morrison C, Jacob ST. Protein tyrosine phosphatase receptor-type O (PTPRO) exhibits characteristics of a candidate tumor suppressor in human lung cancer. *Proc Natl Acad Sci U S A* 2004; 101(38): 13844-9.

Motiwala T, Kutay H, Ghoshal K, Bai S, Seimiya H, Tsuruo T, Suster S, Morrison C, Jacob ST.Protein tyrosine phosphatase receptor-type O (PTPRO) exhibits characteristics of a candidate tumor suppressor in human lung cancer. *Proc Natl Acad Sci U S A* 2004; 101(38): 13844-9.

Müller GA, Engeland K. The central role of CDE/CHR promoter elements in the regulation of cell cycle-dependent gene transcription. *FEBS J* 2010; 277(4): 877-93.

Mumm S, Zhang X, Vacca M, D'Esposito M, Whyte MP. The sedlin gene for spondyloepiphyseal dysplasia tarda escapes X-inactivation and contains a non-canonical splice site. Gene 2001; 273(2): 285-93.

Musunuru K, Strong A, Frank-Kamenetsky M, Lee NE, Ahfeldt T, Sachs KV, Li X, Li H, Kuperwasser N, Ruda VM, Pirruccello JP, Muchmore B, Prokunina-Olsson L, Hall JL, Schadt EE, Morales CR, Lund-Katz S, Phillips MC, Wong J, Cantley W, Racie T, Ejebe KG, Orho-Melander M, Melander O, Koteliansky V, Fitzgerald K, Krauss RM, Cowan CA, Kathiresan S, Rader DJ. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* 2010; 466(7307): 714-19.

Nan H, Hutter CM, Lin Y, Jacobs EJ, Ulrich CM, White E, Baron JA, Berndt SI, Brenner H, Butterbach K, Caan BJ, Campbell PT, Carlson CS, Casey G, Chang-Claude J, Chanock SJ, Cotterchio M, Duggan D, Figueiredo JC, Fuchs CS, Giovannucci EL, Gong J, Haile RW, Harrison TA, Hayes RB, Hoffmeister M, Hopper JL, Hudson TJ, Jenkins MA, Jiao S, Lindor NM, Lemire M, Le Marchand L, Newcomb PA, Ogino S, Pflugeisen BM, Potter JD, Qu C, Rosse SA, Rudolph A, Schoen RE, Schumacher FR, Seminara D, Slattery ML, Thibodeau SN, Thomas F, Thornquist M, Warnick GS, Zanke BW, Gauderman WJ, Peters U, Hsu L, Chan AT; CCFR; GECCO. Association of aspirin and NSAID use with risk of colorectal cancer according to genetic variants. *JAMA* 2015; 313(11): 1133-42.

Nardini M, Gnesutta N, Donati G, Gatta R, Forni C, Fossati A, Vonrhein C, Moras D, Romier C, Bolognesi M, Mantovani R. Sequence-Specific Transcription Factor NF-Y Displays Histone-like DNA Binding and H2B-like Ubiquitination. *Cell* 2013; 152(1-2): 132-43.

Närvä E, Rahkonen N, Emani MR, Lund R, Pursiheimo JP, Nästi J, Autio R, Rasool O, Denessiouk K, Lähdesmäki H, Rao A, Lahesmaa R. RNA-binding protein L1TD1 interacts with LIN28 via RNA and is required for human embryonic stem cell self-renewal and cancer cell proliferation. *Stem Cells* 2012; 30(3): 452-60.

Neufert C, Becker C, Neurath MF. An inducible mouse model of colon carcinogenesis for the analysis of sporadic and inflammation-driven tumor progression. *Nat Protoc* 2007; 2(8): 1998-2004.

Ngeow J, Heald B, Rybicki LA, Orloff MS, Chen JL, Liu X, Yerian L, Willis J, Lehtonen HJ, Lehtonen R, Mester JL, Moline J, Burke CA, Church J, Aaltonen LA, Eng C. Prevalence of germline PTEN, BMPR1A, SMAD4, STK11, and ENG mutations in patients with moderate-load colorectal polyps. *Gastroenterology* 2013; 144(7): 1402-9.

Nguyen HH, Takata R, Akamatsu S, Shigemizu D, Tsunoda T, Furihata M, Takahashi A, Kubo M, Kamatani N, Ogawa O, Fujioka T, Nakamura Y, Nakagawa H. IRX4 at 5p15 suppresses prostate cancer growth through the interaction with vitamin D receptor, conferring prostate cancer susceptibility. Hum Mol Genet 2012; 21(9): 2076-85.

Nica AC, Parts L, Glass D, Nisbet J, Barrett A, Sekowska M, Travers M, Potter S, Grundberg E, Small K, Hedman AK, Bataille V, Tzenova Bell J, Surdulescu G, Dimas AS, Ingle C, Nestle FO, di Meglio P, Min JL, Wilk A, Hammond CJ, Hassanali N, Yang TP, Montgomery SB, O'Rahilly S, Lindgren CM, Zondervan KT, Soranzo N, Barroso I, Durbin R, Ahmadi K, Deloukas P, McCarthy MI, Dermitzakis ET, Spector TD; MuTHER Consortium. The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genet* 2011; 7(2): e1002003.

Nick Barker. Adult intestinal stem cells: critical drivers of epithelial homeostasis and regeneration. Nat Rev Mol Cell Biol 2014; 15(1): 19-33.

Nickels S, Truong T, Hein R, Stevens K, Buck K, Behrens S, Eilber U, Schmidt M, Häberle L, Vrieling A, Gaudet M, Figueroa J, Schoof N, Spurdle AB, Rudolph A, Fasching PA, Hopper JL, Makalic E, Schmidt DF, Southey MC, Beckmann MW, Ekici AB, Fletcher O, Gibson L, Silva Idos S, Peto J, Humphreys MK, Wang J, Cordina-Duverger E, Menegaux F, Nordestgaard BG, Bojesen SE, Lanng C, Anton-Culver H, Ziogas A, Bernstein L, Clarke CA, Brenner H, Müller H, Arndt V, Stegmaier C, Brauch H, Brüning T, Harth V; Genica Network, Mannermaa A, Kataja V, Kosma VM, Hartikainen JM; kConFab; AOCS Management Group, Lambrechts D, Smeets D, Neven P, Paridaens R, Flesch-Janys D, Obi

N, Wang-Gohrke S, Couch FJ, Olson JE, Vachon CM, Giles GG, Severi G, Baglietto L, Offit K, John EM, Miron A, Andrulis IL, Knight JA, Glendon G, Mulligan AM, Chanock SJ, Lissowska J, Liu J, Cox A, Cramp H, Connley D, Balasubramanian S, Dunning AM, Shah M, Trentham-Dietz A, Newcomb P, Titus L, Egan K, Cahoon EK, Rajaraman P, Sigurdson AJ, Doody MM, Guénel P, Pharoah PD, Schmidt MK, Hall P, Easton DF, Garcia-Closas M, Milne RL, Chang-Claude J. Evidence of gene-environment interactions between common breast cancer susceptibility loci and established environmental risk factors. *PLoS Genet* 2013; 9(3): e1003284.

Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet* 2010; 6(4): e1000888.

Nicolaides NC, Papadopoulos N, Liu B, Wei YF, Carter KC, Ruben SM, Rosen CA, Haseltine WA, Fleischmann RD, Fraser CM, Adams MD, Venter JC, Dunlop MG, Hamilton SR, Petersen GM, Chapelle ADL, Vogelstein B, Kinzler KW. Mutations of two PMS homologues in hereditary nonpolyposis colon cancer. *Nature* 1994; 371(6492): 75-80.

Nie Z, Du MQ, McAllister-Lucas LM, Lucas PC, Bailey NG, Hogaboam CM, Lim MS, Elenitoba-Johnson KS. Conversion of the LIMA1 tumour suppressor into an oncogenic LMO-like protein by API2-MALT1 in MALT lymphoma. Nat Commun 2015; 6: 5908.

Norat T, Lukanova A, Ferrari P, Riboli E. Meat consumption and colorectal cancer risk: dose-response meta-analysis of epidemiological studies. *Int J Cancer* 2002; 98: p241-56.

Novelli M, Cossu A, Oukrif D, Quaglia A, Lakhani S, Poulsom R, Sasieni P, Carta P, Contini M, Pasca A, Palmieri G, Bodmer W, Tanda F, Wright N. X-inactivation patch size in human female tissue confounds the assessment of tumor clonality. *Proc Natl Acad Sci U S A* 2003; 100(6): 3311-4.

Ordóñez-Morán P, Larriba MJ, Pálmer HG, Valero RA, Barbáchano A, Duñach M, de Herreros AG, Villalobos C, Berciano MT, Lafarga M, Muñoz A. RhoA-ROCK and p38MAPK-MSK1 mediate vitamin D effects on gene expression, phenotype, and Wnt pathway in colon cancer cells. *J Cell Biol* 2008; 183(4): 697-710.

Pai AA, Cain CE, Mizrahi-Man O, De Leon S, Lewellen N, Veyrieras JB, Degner JF, Gaffney DJ, Pickrell JK, Stephens M, Pritchard JK, Gilad Y. The contribution of RNA decay quantitative trait loci to inter-individual variation in steady-state gene expression levels. *PLoS Genet* 2012; 8: e1003000.

Palles C, Cazier JB, Howarth KM, Domingo E, Jones AM, Broderick P, Kemp Z, Spain SL, Guarino E, Salguero I, Sherborne A, Chubb D, Carvajal-Carmona LG, Ma Y, Kaur K, Dobbins S, Barclay E, Gorman M, Martin L, Kovac MB, Humphray S; CORGI Consortium; WGS500 Consortium, Lucassen A, Holmes CC, Bentley D, Donnelly P, Taylor J, Petridis C, Roylance R, Sawyer EJ, Kerr DJ, Clark S, Grimes J, Kearsey SE, Thomas HJ, McVean G, Houlston RS, Tomlinson I. Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. *Nat Genet* 2013; 45(2): 136-44.

Pálmer HG, González-Sancho JM, Espada J, Berciano MT, Puig I, Baulida J, Quintanilla M, Cano A, de Herreros AG, Lafarga M, Muñoz A. Vitamin D(3) promotes the differentiation of colon carcinoma cells by the induction of E-cadherin and the inhibition of beta-catenin signaling. *J Cell Biol* 2001; 154(2): 369-87.

Pálmer HG, González-Sancho JM, Espada J, Berciano MT, Puig I, Baulida J, Quintanilla M, Cano A, de Herreros AG, Lafarga M, Muñoz A. Vitamin D(3) promotes the differentiation of colon carcinoma cells by the induction of E-cadherin and the inhibition of beta-catenin signaling. *J Cell Biol* 2001; 154(2): 369-87.

Palmer M, Prediger E. Assessing RNA Quality. *Ambion TechNotes* 2004; 11(1). Accessible at http://www.ambion.com/techlib/tn/111/8.html

Park K, Woo M, Nam J, Kim JC. Start codon polymorphisms in the vitamin D receptor and colorectal cancer risk. *Cancer Lett* 2006 Jun 18;237(2):199-206.

Pavek P, Pospechova K, Svecova L, Syrova Z, Stejskalova L, Blazkova J, Dvorak Z, Blahos J. Intestinal cell-specific vitamin D receptor (VDR)-mediated transcriptional regulation of CYP3A4 gene. *Biochem Pharmacol* 2010; 79(2): 277-87.

Peltekova VD, Lemire M, Qazi AM, Zaidi SH, Trinh QM, Bielecki R, Rogers M, Hodgson L, Wang M, D'Souza DJ, Zandi S, Chong T, Kwan JY, Kozak K, De Borja R, Timms L, Rangrej J, Volar M, Chan-Seng-Yue M, Beck T, Ash C, Lee S, Wang J, Boutros PC, Stein LD, Dick JE, Gryfe R, McPherson JD, Zanke BW, Pollett A, Gallinger S, Hudson TJ. Identification of genes expressed by immune cells of the colon that are regulated by colorectal cancer-associated variants. *Int J Cancer* 2014; 134(10): 2330-41.

Peña C, García JM, Silva J, García V, Rodríguez R, Alonso I, Millán I, Salas C, de Herreros AG, Muñoz A, Bonilla F. E-cadherin and vitamin D receptor regulation by SNAIL and ZEB1 in colon cancer: clinicopathological correlations. *Hum Mol Genet* 2005; 14(22): 3361-70.

Peña C, García JM, Silva J, García V, Rodríguez R, Alonso I, Millán I, Salas C, de Herreros AG, Muñoz A, Bonilla F. E-cadherin and vitamin D receptor regulation by SNAIL and ZEB1 in colon cancer: clinicopathological correlations. *Hum Mol Genet* 2005; 14(22): 3361-70.

Peterlik M, Cross HS. Vitamin D and calcium deficits predispose for multiple chronic diseases. *Eur J Clin Invest* 2005; 35(5): 290-304.

Peters U, Jiao S, Schumacher FR, Hutter CM, Aragaki AK, Baron JA, Berndt SI, Bézieau S, Brenner H, Butterbach K, Caan BJ, Campbell PT, Carlson CS, Casey G, Chan AT, Chang-Claude J, Chanock SJ, Chen LS, Coetzee GA, Coetzee SG, Conti DV, Curtis KR, Duggan D, Edwards T, Fuchs CS, Gallinger S, Giovannucci EL, Gogarten SM, Gruber SB, Haile RW, Harrison TA, Hayes RB, Henderson BE, Hoffmeister M, Hopper JL, Hudson TJ, Hunter DJ, Jackson RD, Jee SH, Jenkins MA, Jia WH, Kolonel LN, Kooperberg C, Küry S, Lacroix AZ, Laurie CC, Laurie CA, Le Marchand L, Lemire M, Levine D, Lindor NM, Liu Y, Ma J, Makar KW, Matsuo K, Newcomb PA, Potter JD, Prentice RL, Qu C, Rohan T, Rosse SA, Schoen RE, Seminara D, Shrubsole M, Shu XO, Slattery ML, Taverna D, Thibodeau SN, Ulrich CM, White E, Xiang Y, Zanke BW, Zeng YX, Zhang B, Zheng W, Hsu L; Colon Cancer Family Registry and the Genetics and Epidemiology of Colorectal Cancer Consortium. Identification of Genetic Susceptibility Loci for Colorectal Tumors in a Genome-Wide Meta-analysis. *Gastroenterology* 2013; 144(4): 799-807.

Pfaffl MW, Tichopad A, Prgomet C, Neuvians TP. Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: BestKeeper–Excel-based tool using pair-wise correlations. *Biotechnol Lett* 2004; 26(6): 509-15.

Pike JW, Meyer MB. The Vitamin D Receptor: New Paradigms for the Regulation of Gene Expression by 1,25-Dihydroxyvitamin $D_{3.}$ *Endocrinol Metab Clin North Am* 2010; 39(2): 255-69.

Pinchuk IV, Mifflin RC, Saada JI, Powell DW. Intestinal Mesenchymal Cells. *Curr Gastroenterol Rep* 2010; 12(5): 310-8.

Pino MS, Chung DC. The chromosomal instability pathway in colon cancer. *Gastroenterology* 2010;138(6): 2059-72.

Pinto D, Gregorieff A, Begthel H, Clevers H. Canonical Wnt signals are essential for homeostasis of the intestinal epithelium. *Genes Dev* 2003; 17(14): 1709-13.

Pittman AM, Naranjo S, Jalava SE, Twiss P, Ma Y, Olver B, Lloyd A, Vijayakrishnan J, Qureshi M, Broderick P, van Wezel T, Morreau H, Tuupanen S, Aaltonen LA, Alonso ME, Manzanares M, Gavilán A, Visakorpi T, Gómez-Skarmeta JL, Houlston RS. Allelic Variation at the 8q23.3 Colorectal Cancer Risk Locus Functions as a Cis-Acting Regulator of EIF3H. *PLoS Genet* 2010; 6(9): e1001126.

Pittman AM, Naranjo S, Webb E, Broderick P, Lips EH, van Wezel T, Morreau H, Sullivan K, Fielding S, Twiss P, Vijayakrishnan J, Casares F, Qureshi M, Gómez-Skarmeta JL, Houlston RS.The colorectal cancer risk at 18q21 is caused by a novel variant altering smad7 expression. *Genome Res* 2009; 19(6): 987-93.

Pittman AM, Webb E, Carvajal-Carmona L, Howarth K, Di Bernardo MC, Broderick P, Spain S, Walther A, Price A, Sullivan K, Twiss P, Fielding S, Rowan A, Jaeger E, Vijayakrishnan J, Chandler I, Penegar S, Qureshi M, Lubbe S, Domingo E, Kemp Z, Barclay E, Wood W, Martin L, Gorman M, Thomas H, Peto J, Bishop T, Gray R, Maher ER, Lucassen A, Kerr D, Evans GR; CORGI Consortium, van Wezel T, Morreau H, Wijnen JT, Hopper JL, Southey MC, Giles GG, Severi G, Castellví-Bel S, Ruiz-Ponte C, Carracedo A, Castells A; EPICOLON Consortium, Försti A, Hemminki K, Vodicka P, Naccarati A, Lipton L, Ho JW, Cheng KK, Sham PC, Luk J, Agúndez JA, Ladero JM, de la Hoya M, Caldés T, Niittymäki I, Tuupanen S, Karhu A, Aaltonen LA, Cazier JB, Tomlinson IP, Houlston RS. Refinement of the basis and impact of common 11q23.1 variation to the risk of developing colorectal cancer. *Hum Mol Genet* 2008; 17(23): 3720-7.

Popovici V, Goldstein DR, Antonov J, Jaggi R, Delorenzi M, Wirapati P. Selecting control genes for RT-QPCR using public microarray data. *BMC Bioinformatics* 2009; 10: 42.

Powell DW, Pinchuk IV, Saada JI, Chen X, Mifflin RC. Mesenchymal Cells of the Intestinal Lamina Propria. *Annu Rev Physiol* 2011; 73: 213-7.

Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, Boehnke M, Abecasis GR, Willer CJ. LocusZoom: Regional visualization of genome-wide association scan results. *Bioinformatics* 2010; 26(18): 2336-7.

Rabeneck L, Davila JA, El-Serag HB. Is there a true 'shift' to the right colon in the incidence of colorectal cancer? *Am J Gastroenterol 2003;* 98: p1400-9.

Rajeevan MS, Vernon SD, Taysavang N, Unger ER. Validation of arraybased gene expression profiles by real-time (kinetic) RT-PCR. *J Mol Diagn* 2001; 3(1), 26-31.

Ramagopalan SV, Heger A, Berlanga AJ, Maugeri NJ, Lincoln MR, Burrell A, Handunnetthi L, Handel AE, Disanto G, Orton SM, Watson CT, Morahan JM, Giovannoni G, Ponting CP, Ebers GC, Knight JC. A ChIP-seq defined genome-wide map of vitamin D receptor binding: associations with disease and evolution. *Genome Res* 2010; 20(10): 1352-60.

Raskov H, Pommergaard H-C, Burcharth J, Rosenberg J. Colorectal carcinogenesis-update and perspectives. *World J Gastroenterol* 2014; 20(48): 18151-64.

Regula J, Rupinski M, Kraszewska E, Polkowski M, Pachlewski J, Orlowska J, Nowacki MP, Butruk E. Colonoscopy in colorectal-cancer screening for detection of advanced neoplasia. *N Engl J Med* 2006; 355(18): 1863-72.

Reich DE, Lander ES. On the allelic spectrum of human disease. *Trends Genet* 2001; 17: 502-10.

Reid D, Toole BJ, Knox S, Talwar D, Harten J, O'Reilly DS, Blackwell S, Kinsella J, McMillan DC, Wallace AM. The relation between acute changes in the systemic inflammatory response and plasma 25-hydroxyvitamin D concentrations after elective knee arthroplasty. *Am J Clin Nutr* 2011; 93(5): 1006-11.

Reynolds A, Wharton N, Parris A, Mitchell E, Sobolewski A, Kam C, Bigwood L, El Hadi A, Münsterberg A, Lewis M, Speakman C, Stebbings W, Wharton R, Sargen K, Tighe R, Jamieson C, Hernon J, Kapur S, Oue N, Yasui W, Williams MR. Canonical Wnt signals combined with suppressed TGFβ/BMP pathways promote renewal of the native human colonic epithelium. *Gut* 2014; 63(4): 610-21.

Rinn JL, Snyder M. Sexual dimorphism in mammalian gene expression. *Trends Genet* 2005; 21(5): 298-305.

Roberts JA, Waters L, Ro JY, Zhai QJ. Smoothelin and caldesmon are reliable markers for distinguishing muscularis propria from desmoplasia: a critical distinction for accurate staging colorectal adenocarcinoma. *Int J Clin Exp Pathol* 2014; 7(2): 792-6.

Robsahm TE, Tretli S, Dahlback A, Moan J. Vitamin D3 from sunlight may improve the prognosis of breast-, colon- and prostate cancer (Norway). *Cancer Causes Control* 2004; 15(2): 149-58.

Rockman MV, Kruglyak L. Genetics of global gene expression. *Nature Rev Genet* 2006; 7(11): 862-72.

Rothenberg ME, Nusse Y, Kalisky T, Lee JJ, Dalerba P, Scheeren F, Lobo N, Kulkarni S, Sim S, Qian D, Beachy PA, Pasricha PJ, Quake SR, Clarke MF. Identification of a cKit$^+$ colonic crypt base secretory cell that supports Lgr5$^+$ stem cells in mice. *Gastroenterology* 2012; 142(5): 1195-205.

Rougeulle C, Navarro P, Avner P. Promoter-restricted H3 Lys 4 di-methylation is an epigenetic mark for monoallelic expression. *Hum Mol Genet* 2003; 12(24): 3343-8.

Rustgi A. The genetics of hereditary colon cancer. *Genes Dev* 2007; 21: p2525-38.

Sadanandam A, Lyssiotis CA, Homicsko K, Collisson EA, Gibb WJ, Wullschleger S, Ostos LC, Lannon WA, Grotzinger C, Del Rio M, Lhermitte B, Olshen AB, Wiedenmann B, Cantley LC, Gray JW, Hanahan D. A colorectal cancer classification system that associates cellular phenotype and responses to therapy. *Nat Med* 2013; 19(5): 619-25.

Salazar R, Roepman P, Capella G, Moreno V, Simon I, Dreezen C, Lopez-Doriga A, Santos C, Marijnen C, Westerga J, Bruin S, Kerr D, Kuppen P, van de Velde C, Morreau H, Van Velthuysen L, Glas AM, Van't Veer LJ, Tollenaar R. Gene expression signature to improve prognosis prediction of stage II and III colorectal cancer. *J Clin Oncol* 2011; 29(1): 17-24.

Salsi V, Caretti G, Wasner M, Reinhard W, Haugwitz U, Engeland K, Mantovani R. Interactions between p300 and multiple NF-Y trimers govern cyclin B2 promoter function. *J Biol Chem* 2003; 278(9): 6642-50.

Sangiorgi E, Capecchi MR. Bmi1 is expressed in vivo in intestinal stem cells. *Nat Genet* 2008; 40(7): 915-20.

Sankaran-Walters S, Macal M, Grishina I, Nagy L, Goulart L, Coolidge K, Li J, Fenton A, Williams T, Miller MK, Flamm J, Prindiville T, George M, Dandekar S. Sex differences matter in the gut: effect on mucosal immune activation and inflammation. *Biol Sex Differ* 2013; 4(1): 10.

Sansom OJ, Reed KR, Hayes AJ, Ireland H, Brinkmann H, Newton IP, Batlle E, Simon-Assmann P, Clevers H, Nathke IS, Clarke AR, Winton DJ. Loss of Apc in vivo immediately perturbs Wnt signaling, differentiation, and migration. *Genes Dev* 2004; 18(12): 1385-90.

Sato T, Stange DE, Ferrante M, Vries RG, Van Es JH, Van den Brink S, Van Houdt WJ, Pronk A, Van Gorp J, Siersema PD, Clevers H. Long-term expansion of epithelial organoids from human colon, adenoma, adenocarcinoma, and Barrett's epithelium. *Gastroenterology* 2011; 141(5): 1762-72.

Scanlan PD, Shanahan F, Clune Y, Collins JK, O'Sullivan GC, O'Riordan M, Holmes E, Wang Y, Marchesi JR. Culture-independent analysis of the gut microbiota in colorectal cancer and polyposis. *Environ Microbiol* 2008; 10(3): p789-98.

Schiaffino MV, Baschirotto C, Pellegrini G, Montalti S, Tacchetti C, De Luca M, Ballabio A. The ocular albinism type 1 gene product is a membrane glycoprotein localized to melanosomes. *Proc Natl Acad Sci U S A* 1996; 93(17): 9055-60.

Schoeftner S, Blanco R, Lopez de Silanes I, Muñoz P, Gómez-López G, Flores JM, Blasco MA. Telomere shortening relaxes X chromosome inactivation and forces global transcriptome alterations. *Proc Natl Acad Sci U S A* 2009; 106(46): 19393-8.

Schroeder A, Mueller O, Stocker S, Salowsky R, Leiber M, Gassmann M, Lightfoot S, Menzel W, Granzow M, Ragg T. The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC Mol Biol* 2006; 7: 3

Schwabe RF, Jobin C. The microbiome and cancer. *Nat Rev Cancer* 2013; 13(11): p800-12.

Schwitalla S, Fingerle AA, Cammareri P, Nebelsiek T, Göktuna SI, Ziegler PK, Canli O, Heijmans J, Huels DJ, Moreaux G, Rupec RA, Gerhard M, Schmid R, Barker N, Clevers H, Lang R, Neumann J, Kirchner T, Taketo MM, van den Brink GR, Sansom OJ, Arkan MC, Greten FR. Intestinal tumorigenesis initiated by dedifferentiation and acquisition of stem-cell-like properties. *Cell* 2013; 152(1–2): 25-38.

Sharma SM, Choi D, Planck SR, Harrington CA, Austin CR, Lewis JA, Diebel TN, Martin TM, Smith JR, Rosenbaum JT. Insights in to the pathogenesis of axial spondyloarthropathy based on gene expression profiles. *Arthritis Res Ther* 2009; 11(6): R168.

Shen L, Kondo Y, Rosner GL, Xiao L, Hernandez NS, Vilaythong J, Houlihan PS, Krouse RS, Prasad AR, Einspahr JG, Buckmeier J, Alberts DS, Hamilton SR, Issa JP. MGMT promoter methylation and field defect in sporadic colorectal cancer. *J Natl Cancer Inst* 2005; 97(18): 1330-8.

Shen L, Toyota M, Kondo Y, Lin E, Zhang L, Guo Y, Hernandez NS, Chen X, Ahmed S, Konishi K, Hamilton SR, Issa JP. Integrated genetic and epigenetic analysis identifies three different subclasses of colon cancer. *Proc Natl Acad Sci USA* 2007; 104(47): 18654-59.

Shiota M, Song Y, Yokomizo A, Kiyoshima K, Tada Y, Uchino H, Uchiumi T, Inokuchi J, Oda Y, Kuroiwa K, Tatsugami K, Naito S. Foxo3a suppression of urothelial cancer invasiveness through Twist1, Y-box-binding protein 1, and E-cadherin regulation. *Clin Cancer Res* 2010; 16(23): 5654-63.

Shiratori H, Yashiro K, Shen MM, Hamada H. Conserved regulation and role of Pitx2 in situs-specific morphogenesis of visceral organs. Development 2006; 133(15): 3015-25.

Shou Y, Robinson DM, Amakye DD, Rose KL, Cho YJ, Ligon KL, Sharp T, Haider AS, Bandaru R, Ando Y, Geoerger B, Doz F, Ashley DM, Hargrave DR, Casanova M, Tawbi HA, Rodon J, Thomas AL, Mita AC, MacDonald TJ, Kieran MW. A five-gene hedgehog signature developed as a patient preselection tool for hedgehog inhibitor therapy in medulloblastoma. *Clin Cancer Res* 2015; 21(3): 585-93.

Sieber OM, Lipton L, Crabtree M, Heinimann K, Fidalgo P, Phillips RK, Bisgaard ML, Orntoft TF, Aaltonen LA, Hodgson SV, Thomas HJ, Tomlinson IP. Multiple colorectal adenomas, classic adenomatous polyposis, and germ-line mutations in MYH. *N Engl J Med* 2003; 348(9): p791-9.

Silver N, Best S, Jiang J, Thein SL. Selection of housekeeping genes for gene expression studies in human reticulocytes using real-time PCR. *BMC Mol Biol* 2006; 7: 33.

Smith PM, Howitt MR, Panikov N, Michaud M, Gallini CA, Bohlooly-Y M, Glickman JN, Garrett WS. The microbial metabolites, short-chain fatty acids, regulate colonic Treg cell homeostasis. *Science* 2013; 341(6145): p569-73.

Somel M, Khaitovich P, Bahn S, Pääbo S, Lachmann M. Gene expression becomes heterogeneous with age. *Curr Biol* 2006; 16(10): R359-60.

Somsen D, Davis-Keppen L, Crotwell P, Flanagan J, Munson P, Stein Q. Congenital nasal pyriform aperture stenosis and ocular albinism co-occurring in a sibship with a maternally-inherited 97 kb Xp22.2 microdeletion. *Am J Med Genet* 2014; 164A(5):1268-71.

Sotelo J, Esposito D, Duhagon MA, Banfield K, Mehalko J, Liao H, Stephens RM, Harris TJ, Munroe DJ, Wu X. Long-range enhancers on 8q24 regulate c-Myc. *Proc Natl Acad Sci U S A* 2010; 107(7): 3001-5.

Spruessel A, Steimann G, Jung M, Lee SA, Carr T, Fentz AK, Spangenberg J, Zornig C, Juhl HH, David KA. Tissue ischemia time affects gene and protein expression patterns within minutes following surgical tumor excision. *Biotechniques* 2004, 36(6): 1030-7.

Stemmler MP, Hecht A, Kemler R. E-cadherin intron 2 contains cis-regulatory elements essential for gene expression. *Development* 2005; 132(5): 965-76.

Strand C, Enell J, Hedenfalk I, Fernö M. RNA quality in frozen breast cancer samples and the influence on gene expression analysis--a comparison of three evaluation methods using microcapillary electrophoresis traces. *BMC Mol Biol* 2007; 8: 38.

Stranger BE, Forrest MS, Clark AG, Minichiello MJ, Deutsch S, Lyle R, Hunt S, Kahl B, Antonarakis SE, Tavaré S, Deloukas P, Dermitzakis ET. Genome-wide associations of gene expression variation in humans. *PLoS Genet* 2005;1(6): e78.

Subbaiah VK, Kranjec C, Thomas M, Banks L. PDZ domains: the building blocks regulating tumorigenesis. *Biochem J* 2011; 439(2): 195-205.

Sung JJ, Lau JY, Goh KL, Leung WK; Asia Pacific Working Group on Colorectal Cancer. Increasing incidence of colorectal cancer in Asia: implications for screening. *Lancet Oncol* 2005, 6 (11): p871-6.

Sutton AL, MacDonald PN. Vitamin D: more than a "bone-a-fide" hormone. *Mol Endocrinol* 2003; 17(5): 777-91.

Tangrea J, Helzlsouer K, Pietinen P, Taylor P, Hollis B, Virtamo J, Albanes D. Serum levels of vitamin D metabolites and the subsequent risk of colon and rectal cancer in Finnish men. *Cancer Causes Control* 1997; 8(4): 615-25.

Tenesa A, Farrington SM, Prendergast JG, Porteous ME, Walker M, Haq N, Barnetson RA, Theodoratou E, Cetnarskyj R, Cartwright N, Semple C, Clark AJ, Reid FJ, Smith LA, Kavoussanakis K, Koessler T, Pharoah PD, Buch S, Schafmayer C, Tepel J, Schreiber S, Völzke H, Schmidt CO, Hampe J, Chang-Claude J, Hoffmeister M, Brenner H, Wilkening S, Canzian F, Capella G, Moreno V, Deary IJ, Starr JM, Tomlinson IP, Kemp Z, Howarth K, Carvajal-Carmona L, Webb E, Broderick P, Vijayakrishnan J, Houlston RS, Rennert G, Ballinger D, Rozek L, Gruber SB, Matsuda K, Kidokoro T, Nakamura Y, Zanke BW, Greenwood CM, Rangrej J, Kustra R, Montpetit A, Hudson TJ, Gallinger S, Campbell H, Dunlop MG.Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. *Nat Genet 2008;* 40(5): 631-37.

Terzić J, Grivennikov S, Karin E, Karin M. Inflammation and colon cancer. *Gastroenterology* 2010; 138(6): 2101-14

Testa A, Donati G, Yan P, Romani F, Huang TH, Viganò MA, Mantovani R. Chromatin immunoprecipitation (ChIP) on chip experiments uncover a widespread distribution of NF-Y binding CCAAT sites outside of core promoters. *J Biol Chem* 2005; 280(14): 13606-15.

Tetsu O, McCormick F. Beta-catenin regulates expression of cyclin D1 in colon carcinoma cells. *Nature* 1999; 398(6726): 422-6.

Tomlinson I, Webb E, Carvajal-Carmona L, Broderick P, Kemp Z, Spain S, Penegar S, Chandler I, Gorman M, Wood W, Barclay E, Lubbe S, Martin L, Sellick G, Jaeger E, Hubner R, Wild R, Rowan A, Fielding S, Howarth K; CORGI Consortium, Silver A, Atkin

W, Muir K, Logan R, Kerr D, Johnstone E, Sieber O, Gray R, Thomas H, Peto J, Cazier JB, Houlston R. A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nat Genet 2007;* 39(8), 984-88.

Tomlinson IP, Carvajal-Carmona LG, Dobbins SE, Tenesa A, Jones AM, Howarth K, Palles C, Broderick P, Jaeger EE, Farrington S, Lewis A, Prendergast JG, Pittman AM, Theodoratou E, Olver B, Walker M, Penegar S, Barclay E, Whiffin N, Martin L, Ballereau S, Lloyd A, Gorman M, Lubbe S; COGENT Consortium; CORGI Collaborators; EPICOLON Consortium, Howie B, Marchini J, Ruiz-Ponte C, Fernandez-Rozadilla C, Castells A, Carracedo A, Castellvi-Bel S, Duggan D, Conti D, Cazier JB, Campbell H, Sieber O, Lipton L, Gibbs P, Martin NG, Montgomery GW, Young J, Baird PN, Gallinger S, Newcomb P, Hopper J, Jenkins MA, Aaltonen LA, Kerr DJ, Cheadle J, Pharoah P, Casey G, Houlston RS, Dunlop MG. Multiple common susceptibility variants near BMP pathway loci GREM1, BMP4, and BMP2 explain part of the missing heritability of colorectal cancer. *PLoS Genet* 2011; 7(6): e1002105.

Tomlinson IP, Webb E, Carvajal-Carmona L, Broderick P, Howarth K, Pittman AM, Spain S, Lubbe S, Walther A, Sullivan K, Jaeger E, Fielding S, Rowan A, Vijayakrishnan J, Domingo E, Chandler I, Kemp Z, Qureshi M, Farrington SM, Tenesa A, Prendergast JG, Barnetson RA, Penegar S, Barclay E, Wood W, Martin L, Gorman M, Thomas H, Peto J, Bishop DT, Gray R, Maher ER, Lucassen A, Kerr D, Evans DG; CORGI Consortium, Schafmayer C, Buch S, Völzke H, Hampe J, Schreiber S, John U, Koessler T, Pharoah P, van Wezel T, Morreau H, Wijnen JT, Hopper JL, Southey MC, Giles GG, Severi G, Castellví-Bel S, Ruiz-Ponte C, Carracedo A, Castells A; EPICOLON Consortium, Försti A, Hemminki K, Vodicka P, Naccarati A, Lipton L, Ho JW, Cheng KK, Sham PC, Luk J, Agúndez JA, Ladero JM, de la Hoya M, Caldés T, Niittymäki I, Tuupanen S, Karhu A, Aaltonen L, Cazier JB, Campbell H, Dunlop MG, Houlston RS. A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. *Nat Genet* 2008; 40(5): 623-30.

Touvier M, Chan DS, Lau R, Aune D, Vieira R, Greenwood DC, Kampman E, Riboli E, Hercberg S, Norat T. Meta-analyses of vitamin D intake, 25-hydroxyvitamin D status, vitamin D receptor polymorphisms, and colorectal cancer risk. *Cancer Epidemiol Biomarkers Prev* 2011 May; 20(5): 1003-16.

Towler BP, Irwig L, Glasziou P, Weller D, Kewenter J. Screening for colorectal cancer using the faecal occult blood test, Hemoccult. *Cochrane Database Syst Rev* 2007; (1): CD001216.

Tretli S, Schwartz GG, Torjesen PA, Robsahm TE. Serum levels of 25-hydroxyvitamin D and survival in Norwegian patients with cancer of breast, colon, lung, and lymphoma: a population-based study. *Cancer Causes Control* 2012; 23(2): 363-70.

Trivedi DP, Doll R, Khaw KT. Effect of four monthly oral vitamin D3 (cholecalciferol) supplementation on fractures and mortality in men and women living in the community: randomized double blind controlled trial. *BMJ* 2003; 326(7387): 469.

Tsanou E, Peschos D, Batistatou A, Charalabopoulos A, Charalabopoulos K. The E-cadherin adhesion molecule and colorectal cancer. A global literature approach. *Anticancer Res*. 2008 Nov-Dec;28(6A):3815-26.

Uitterlinden A, Fang Y, Van Meurs J, Pols H, Van Leeuwen J. Genetics and biology of vitamin D receptor polymorphisms. *Gene* 2004; 338(2): 143–56.

van de Wetering M, Sancho E, Verweij C, de Lau W, Oving I, Hurlstone A, van der Horn K, Batlle E, Coudreuse D, Haramis AP, Tjon-Pon-Fong M, Moerer P, van den Born M, Soete G, Pals S, Eilers M, Medema R, Clevers H. The beta-Catenin/TCF-4 Complex Imposes a Crypt Progenitor Phenotype on Colorectal Cancer Cells. *Cell* 2002; 111(2): 241-50.

Van den Bossche J, Malissen B, Mantovani A, De Baetselier P, Van Ginderachter JA. Regulation and function of the E-cadherin/catenin complex in cells of the monocyte-macrophage lineage and DCs. *Blood* 2012; 119(7): 1623-33.

Van den Bossche J, Van Ginderachter JA. E-cadherin: from epithelial glue to immunological regulator. *Eur J Immunol* 2013; 43(1): 34-7.

van der Flier LG, Haegebarth A, Stange DE, van de Wetering M, Clevers H. OLFM4 is a robust marker for stem cells in human intestine and marks a subset of colorectal cancer cells. *Gastroenterology* 2009; 137(1): 15-7.

van Es JH, van Gijn ME, Riccio O, van den Born M, Vooijs M, Begthel H, Cozijnsen M, Robine S, Winton DJ, Radtke F, Clevers H. Notch/γ-secretase inhibition turns proliferative cells in intestinal crypts and adenomas into goblet cells. *Nature* 2005; 435(7044): 959-63.

Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A, Speleman F. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol* 2002; 3(7): RESEARCH0034.

Vandesompele J, Kubista M, Pfaffl MW. Reference gene validation software for improved normalization. In Logan J, Edwards Kirstin, Saunders N (eds). *Real-Time PCR: Current Technology and Applications.* Caister Academic Press; 2009. p47-64.

Veigl ML, Kasturi L, Olechnowicz J, Ma AH, Lutterbaugh JD, Periyasamy S, Li GM, Drummond J, Modrich PL, Sedwick WD, Markowitz SD. Biallelic inactivation of hMLH1 by epigenetic gene silencing, a novel mechanism causing human MSI cancers. *Proc Natl Acad Sci USA* 1998; 95(15): 8698-702.

Visvader J. Cells of origin in cancer. *Nature* 2011; 469(7330): 314-22.

von Holst S, Picelli S, Edler D, Lenander C, Dalén J, Hjern F, Lundqvist N, Lindforss U, Påhlman L, Smedh K, Törnqvist A, Holm J, Janson M, Andersson M, Ekelund S, Olsson L, Ghazi S, Papadogiannakis N, Tenesa A, Farrington SM, Campbell H, Dunlop MG, Lindblom A. Association studies on 11 published colorectal cancer risk loci. *Br J Cancer* 2010; 103(4): 575-80.

Wallace AM, Gibson S, de la Hunty A, Lamberg-Allardt C, Ashwell M. Measurement of 25-hydroxyvitamin D in the clinical laboratory: current procedures, performance characteristics and limitations. *Steroids* 2010; 75(7): 477-88.

Wang J, Zhuang J, Iyer S, Lin X, Whitfield TW, Greven MC, Pierce BG, Dong X, Kundaje A, Cheng Y, Rando OJ, Birney E, Myers RM, Noble WS, Snyder M, Weng Z. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res* 2012; 22(9): 1798-812

Wang J, Zhuang J, Iyer S, Lin XY, Greven MC, Kim BH, Moore J, Pierce BG, Dong X, Virgil D, Birney E, Hung JH, Weng Z. Factorbook.org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium. *Nucleic Acids Res* 2013; 41(Database issue): D171-6.

Wang T, Cai G, Qiu Y, Fei N, Zhang M, Pang X, Jia W, Cai S, Zhao L. Structural segregation of gut microbiota between colorectal cancer patients and healthy volunteers. *ISME J* 2012; 6(2): p320-9.

Wang Y, Barbacioru C, Hyland F, Xiao W, Hunkapiller KL, Blake J, Chan F, Gonzalez C, Zhang L, Samaha RR. Large scale real-time PCR validation on gene expression measurements from two commercial long-oligonucleotide microarrays. *BMC Genomics* 2006; 7: 59.

Watanabe T, Wang S, Noritake J, Sato K, Fukata M, Takefuji M, Nakagawa M, Izumi N, Akiyama T, Kaibuchi K. Interaction with IQGAP1 Links APC to Rac1, Cdc42, and Actin Filaments during Cell Polarization and Migration. *Dev Cell* 2004; 7(6): 871-83.

Weisenberger DJ, Siegmund KD, Campan M, Young J, Long TI, Faasse MA, Kang GH, Widschwendter M, Weener D, Buchanan D, Koh H, Simms L, Barker M, Leggett B, Levine J, Kim M, French AJ, Thibodeau SN, Jass J, Haile R, Laird PW. CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer. *Nat Genet* 2006; 38(7): 787-793.

Westphalen CB, Asfaha S, Hayakawa Y, Takemoto Y, Lukin DJ, Nuber AH, Brandtner A, Setlik W, Remotti H, Muley A, Chen X, May R, Houchen CW, Fox JG, Gershon MD, Quante M, Wang TC. Long-lived intestinal tuft cells serve as colon cancer-initiating cells. *J Clin Invest 2014;* 124(3): 1283-95.

Whiffin N, Dobbins SE, Hosking FJ, Palles C, Tenesa A, Wang Y, Farrington SM, Jones AM, Broderick P, Campbell H, Newcomb PA, Casey G, Conti DV, Schumacher F, Gallinger S, Lindor NM, Hopper J, Jenkins M, Dunlop MG, Tomlinson IP, Houlston RS. Deciphering the genetic architecture of low-penetrance susceptibility to colorectal cancer. *Hum Mol Genet* 2013; 22(24): 5075-82.

Whiffin N, Hosking FJ, Farrington SM, Palles C, Dobbins SE, Zgaga L, Lloyd A, Kinnersley B, Gorman M, Tenesa A, Broderick P, Wang Y, Barclay E, Hayward C, Martin L, Buchanan DD, Win AK, Hopper J, Jenkins M, Lindor NM, Newcomb PA, Gallinger S, Conti D, Schumacher F, Casey G, Liu T; Swedish Low-Risk Colorectal Cancer Study Group, Campbell H, Lindblom A, Houlston RS, Tomlinson IP, Dunlop MG. Identification of susceptibility loci for colorectal cancer in a genome-wide meta-analysis. *Hum Mol Genet* 2014; 23(17): p4729-37.

Whiffin N, Houlston RS. Architecture of Inherited Susceptibility to Colorectal Cancer: A Voyage of Discovery. *Genes* 2014; *5*(2): 270-84.

Whitfield GK, Remus LS, Jurutka PW, Zitzer H, Oza AK, Dang HT, Haussler CA, Galligan MA, Thatcher ML, Encinas Dominguez C, Haussler MR. Functionally relevant polymorphisms in the human nuclear vitamin D receptor gene. *Mol Cell Endocrinol* 2001;177(1-2):145-59.

Wills JC, Burt RW. Hereditary Aspects of Colon Cancer. *The Ochsner Journal* 2002; 4(3): 129-38.

Wiseman M. The second World Cancer Research Fund/American Institute for Cancer Research expert report. Food, nutrition, physical activity, and the prevention of cancer: a global perspective. *Proc Nutr Soc* 2008; 67(3): p253-6.

Wong HL, Seow A, Arakawa K, Lee HP, Yu MC, Ingles SA. Vitamin D receptor start codon polymorphism and colorectal cancer risk: effect modification by dietary calcium and fat in Singapore Chinese. *Carcinogenesis* 2003;24(6):1091-5.

Wong RC, Ibrahim A, Fong H, Thompson N, Lock LF, Donovan PJ. L1TD1 is a marker for undifferentiated human embryonic stem cells. *PLoS One* 2011; 6(4): e19355.

Wood LD, Parsons DW, Jones S, Lin J, Sjöblom T, Leary RJ, Shen D, Boca SM, Barber T, Ptak J, Silliman N, Szabo S, Dezso Z, Ustyanksky V, Nikolskaya T, Nikolsky Y, Karchin R, Wilson PA, Kaminker JS, Zhang Z, Croshaw R, Willis J, Dawson D, Shipitsin M, Willson JK, Sukumar S, Polyak K, Park BH, Pethiyagoda CL, Pant PV, Ballinger DG, Sparks AB, Hartigan J, Smith DR, Suh E, Papadopoulos N, Buckhaults P, Markowitz SD, Parmigiani G, Kinzler KW, Velculescu VE, Vogelstein B.The genomic landscapes of human breast and colorectal cancers. *Science* 2007; 318(5853): 1108-13.

Worthley DL, Walsh MD, Barker M, Ruszkiewicz A, Bennett G, Phillips K, Suthers G. Familial Mutations in PMS2 Can Cause Autosomal Dominant Hereditary Nonpolyposis Colorectal Cancer. *Gastroenterology* 2005; 128(5): 1431-6.

Yamamoto Y, Yin MJ, Lin KM, Gaynor RB. Sulindac inhibits activation of the NF-kappaB pathway. *J Biol Chem* 1999; 274(38): p27307-14.

Yang G, Scherer SJ, Shell SS, Yang K, Kim M, Lipkin M, Kucherlapati R, Kolodner RD, Edelmann W. Dominant effects of an Msh6 missense mutation on DNA repair and cancer susceptibility. *Cancer Cell* 2004; 6(2): 139-50.

Yen PH, Ellison J, Salido EC, Mohandas T, Shapiro L. Isolation of a new gene from the distal short arm of the human X chromosome that escapes X-inactivation. Hum Mol Genet 1992; 1(1): 47–52.

Zanke BW, Greenwood CM, Rangrej J, Kustra R, Tenesa A, Farrington SM, Prendergast J, Olschwang S, Chiang T, Crowdy E, Ferretti V, Laflamme P, Sundararajan S, Roumy S, Olivier JF, Robidoux F, Sladek R, Montpetit A, Campbell P, Bezieau S, O'Shea AM, Zogopoulos G, Cotterchio M, Newcomb P, McLaughlin J, Younghusband B, Green R, Green J, Porteous ME, Campbell H, Blanche H, Sahbatou M, Tubacher E, Bonaiti-Pellié C, Buecher B, Riboli E, Kury S, Chanock SJ, Potter J, Thomas G, Gallinger S, Hudson TJ, Dunlop MG.Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nat Genet* 2007; 39(8): 989-94.

Zella LA, Kim S, Shevde NK, Pike JW. Enhancers located within two introns of the vitamin D receptor gene mediate transcriptional autoregulation by 1,25-dihydroxyvitamin D3. *Mol Endocrinol* 2006; 20(6): 1231-47.

Zeller T, Wild P, Szymczak S, Rotival M, Schillert A, Castagne R, Maouche S, Germain M, Lackner K, Rossmann H, Eleftheriadis M, Sinning CR, Schnabel RB, Lubos E, Mennerich D, Rust W, Perret C, Proust C, Nicaud V, Loscalzo J, Hübner N, Tregouet D, Münzel T, Ziegler A, Tiret L, Blankenberg S, Cambien F. Genetics and beyond--the transcriptome of human monocytes and disease susceptibility. *PLoS One* 2010; 5: e10693.

Zgaga L, Theodoratou E, Farrington SM, Din FV, Ooi LY, Glodzik D, Johnston S, Tenesa A, Campbell H, Dunlop MG. Plasma vitamin D concentration influences survival outcome after a diagnosis of colorectal cancer. *J Clin Oncol* 2014; 32(23): 2430-9.

Zhang J, Schweers B, Dyer MA. The first knockout mouse model of retinoblastoma. *Cell Cycle* 2004; 3(7): 952-9.

Zhang W, Duan S, Bleibel WK, Wisel SA, Huang RS, Wu X, He L, Clark TA, Chen TX, Schweitzer AC, Blume JE, Dolan ME, Cox NJ. Identification of common genetic variants that account for transcript isoform variation between human populations. *Hum Genet* 2009; 125(1): 81-93.

Zhong H, Beaulaurier J, Lum PY, Molony C, Yang X, Macneil DJ, Weingarth DT, Zhang B, Greenawalt D, Dobrin R, Hao K, Woo S, Fabre-Suver C, Qian S, Tota MR, Keller MP, Kendziorski CM, Yandell BS, Castro V, Attie AD, Kaplan LM, Schadt EE. Liver and adipose expression associated SNPs are enriched for association to type 2 diabetes. *PLoS Genet* 2010; 6: e1000932.

Zhu L, Gibson P, Currle DS, Tong Y, Richardson RJ, Bayazitov IT, Poppleton H, Zakharenko S, Ellison DW, Gilbertson RJ. Prominin 1 marks intestinal stem cells that are susceptible to neoplastic transformation. *Nature 2009;* 457(7229): 603-7.

Zou L, Zhong R, Lou J, Lu X, Wang Q, Yang Y, Xia J, Ke J, Zhang T, Sun Y, Liu L, Cui Y, Xiao H, Chang L, Xia D, Xu H.Replication study in Chinese population and meta-analysissupports association of the 11q23 locus with colorectal cancer. *PLoS One* 2012; 7(9): e45461.

Zuk O, Hechter E, Sunyaev SR, Lander ES. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc Natl Acad Sci U S A* 2012; 109(4): 1193-8.

Zuo T, Wang L, Morrison C, Chang X, Zhang H, Li W, Liu Y, Wang Y, Liu X, Chan MW, Liu JQ, Love R, Liu CG, Godfrey V, Shen R, Huang TH, Yang T, Park BK, Wang CY, Zheng P, Liu Y. FOXP3 is an X-Linked breast cancer suppressor gene and an important repressor of the HER2/ErbB2 oncogene. *Cell* 2007; 129(7), 1275-86.

URL1.1	http://www.wcrf.org/int/cancer-facts-figures/data-specific-cancers/colorectal-cancer-statistics
URL1.2	http://www.ons.gov.uk/ons/rel/cancer-unit/cancer-incidence-and-mortality/2008-2010/stb-cancer-incidence-and-mortality-in-the-united-kindom--2008-2010.html
URL1.3	http://www.cancerresearchuk.org/cancer-info/cancerstats/incidence/risk/
URL1.4	http://www.who.int/gho/ncd/mortality_morbidity/cancer_text/en/
URL1.5	http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/bowel-cancer/mortality#heading-Zero
URL2.1	http://genome.ucsc.edu/

URL2.2	http://david.abcc.ncifcrf.gov/
URL2.3	http://cbl-gorilla.cs.technion.ac.il/
URL4.1	http://www.leonxie.com/referencegene.php
URL4.2	http://www.lifetechnologies.com/uk/en/home/life-science/pcr/real-time-pcr/real-time-pcr-assays/taqman-gene-expression/taqman-endogenous-controls.html
URL4.3	http://www.lifetechnologies.com/order/catalog/product/4396840
URL6.1	http://locuszoom.sph.umich.edu/locuszoom/
URL7.1	http://www.sanger.ac.uk/research/projects/vertebrategenome/havana/#t_ref
URL7.2	http://www.1000genomes.org
URL7.3	http://www.completegenomics.com/public-data/69-Genomes/
URL7.4	http://www.ncbi.nlm.nih.gov/gene/?term=338917
URL8.1	http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/
URL8.2	http://doubling-time.com/compute.php
URL9.1	http://www.ensembl.org/Homo_sapiens/Gene/Summary?db=core;g=ENSG00000111424;r=12:47841537-47943048