The Search for Genetic Variants that Influence the Risk of Colorectal Cancer

Sarah Louise West

University College London

and

Cancer Research UK London Research Institute

PhD Supervisor: Prof. Ian Tomlinson

A thesis submitted for the degree of

Doctor of Philosophy

University College London

November 2010

Declaration

I, Sarah Louise West, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

The main aim of this thesis was to uncover common low penetrance variants that influence susceptibility to colorectal cancer (CRC). This was largely considered in relation to the analysis of the plethora of genetic data from our large genome-wide association study. My work includes fine-mapping of associated loci through additional genotyping, gene screening, and imputation for the prediction of untyped SNPs, which improved the resolution for fine-mapping and facilitated meta-analysis with datasets typed on different arrays. This led to the identification of 14 independent risk loci, while an association analysis of the X chromosome revealed evidence for two additional risk variants. I cover the detection of runs of homozygous SNPs to investigate the relationship between homozygosity and CRC and show that there is no evidence for increased homozygosity in cases in the UK population. I go on to investigate linkage based techniques to perform an analysis of chromosomal regions identical by descent (IBD), which are shared between unrelated cases more often than controls that could harbour risk variants and identified a number of good candidate genes, such as AXIN2 and E2F7, which require further analysis in additional samples. I also search for moderate penetrance susceptibility variants in several families with a dominant-like inheritance and compare identified linkage peaks with the results of a loss of heterozygosity analysis of tumour DNA from family members to identify potential tumour suppressor genes. This analysis identified several promising regions and led to the detection of a *SMAD4* mutation in one family. The associated variants identified in this study provide good evidence that the common-disease common-variant hypothesis holds true, but that this is not the whole story as these variants

account for just 8% of the familial risk. Further research and techniques will be required to uncover the remaining missing heritability.

This thesis is dedicated to my late brother, Philip A. Spain. His strength, perseverance and patience in living with Duchenne’s muscular dystrophy will always be an inspiration to me.

Acknowledgement

First and foremost I thank my primary supervisor, Prof. Ian Tomlinson, for giving me the opportunity to work with him on exciting projects that challenged me and for inspiring me to be a scientist. Thank you for your wisdom, valuable guidance and for expanding the horizons.

I thank Cancer Research UK for funding my PhD and I am also very appreciative to all members of the MPG laboratory for their support and for making my time with them enjoyable and memorable, in particular, Dr Luis Carvajal-Carmona for his academic discussions and advice (and entertaining money saving tips), Dr Zoe Kemp (with fond memories of CORGI corner) and Kimberley Howarth, Dr Emma Jaeger and Angela Jones for their assistance in the lab.

My thanks go to the Bioinformatics and Biostatistics group at the London Research Institute, particularly Aengus Stewart for his support and lively discussion, Gavin Kelly and Stuart Horsey for statistics advice and Probir Chakravarty for advice on pathway analysis.

Special thanks go to my mentor, in statistical genetics Dr Jean-Baptiste Cazier, for giving me the tools, support and encouragement to pursue an analytical focus and for our motivating discussions, I will be ever grateful.

Big thanks go to my friends and family for their ever present support and understanding and the choir of All Saint's, Frindsbury for keeping me singing.

Finally, I wish to express my love and deep gratitude to my husband, Ant, without whom none of this would have been possible and who has supported my academic endeavours and endured the ups and downs of my research with reassuring good grace. Thank you for making my efforts worthwhile and for sharing the journey.

Table of Contents

Table of Figures

List of Tables

Abbreviations

CI Confidence Interval

CNV Copy Number Variant

CORGI Colorectal Tumour Gene Identification Study

CRC Colorectal Cancer

FAP Familial Adenomatous Polyposis

GWA Genome-Wide Association

HWE Hardy Weinberg Equilibrium

HNPCC Hereditary Non-Polyposis Colon Cancer

IBD Identical By Descent

IBS Identical By State

LD Linkage Disequilibrium

LOD Logarithm of Odds

LOH Loss of Heterozygosity

MAF Minor Allele Frequency

OR Odds Ratio

PCA Principal Components Analysis

QC Quality Control

ROH Run Of Homozygosity

SE Standard Error

SD Standard Deviation

SNP Single Nucleotide Polymorphism

TVA Tubulovillous Adenoma

TA Tubular Adenoma

Chapter 1. Introduction

Owing to advances in technology and genetics, the last five years has seen a plethora of genome-wide association (GWA) studies with the hope of discovering common genetic variants that influence an individual's risk of developing a certain disease. The goal is to determine alleles or loci that explain the heritability of diseases above and beyond the known highly penetrant Mendelian conditions that could be used for risk prediction (heritability is the variance in phenotype that could be explained by inherited factors). These loci will increase our knowledge of the genes involved in diseases and thus can provide information about biological pathways and the biomarkers that could be utilised in treatment strategies. There are, of course, many challenges to overcome in achieving this goal. In this introduction, I will discuss these challenges and the progression of techniques, for the detection of predisposition genes, that has made the GWA study feasible and give an overview of the knowledge progression in relation to colorectal cancer (CRC).

The main focus of this thesis is the application of statistical genetic methods to GWA data and additional efforts to search for susceptibility alleles including candidate gene screens, analysis of runs of homozygosity, somatic loss of heterozygosity and linkage analysis.

1.2 Colorectal Cancer

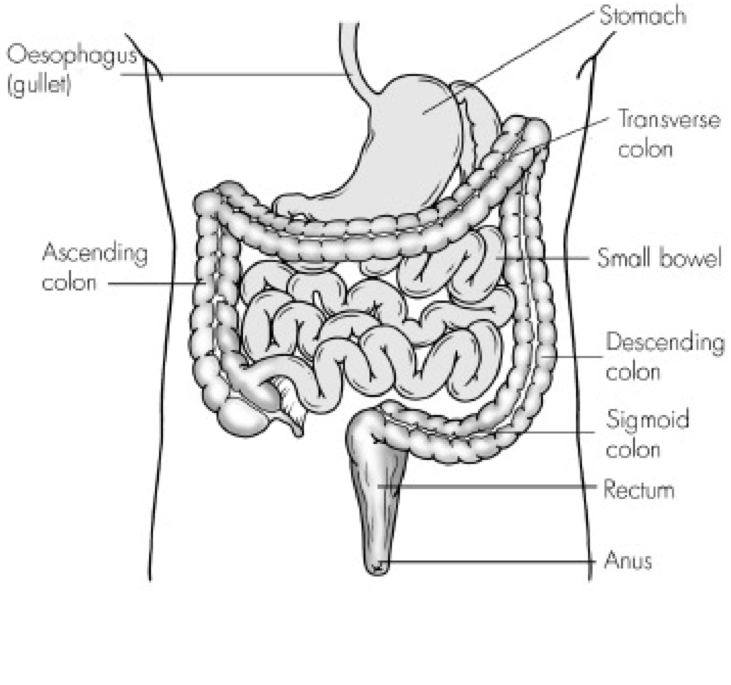
Colorectal Cancer is the third most common cancer in the UK, after breast and lung cancer, with more than 37,500 new cases diagnosed every year and the cause of 16,250 deaths in 2008 (http://info.cancerresearchuk.org/cancerstats/types/bowel/).

The lifetime risk of developing CRC in the general UK population is 5% (Yang *et al.* 2005). However, first degree relatives of individuals with CRC are twice as likely to develop the disease.

CRC can be classified into sporadic (or non-familial) and familial cancers. Familial cancers are caused by an inherited predisposition, such as a germline mutation or polymorphism, which results in familial clustering of cases. Most cancers are thought of as sporadic as they do not cluster in families and can be caused by somatic mutations at the site of the tumour. However, individuals with sporadic cancers may have inherited a low penetrance predisposition to the disease that gives the appearance of a non-familial cancer.

Figure 1.1 Sites affected in CRC

This diagram shows the large colon and rectum, which are the sites of CRC. Approximately, two thirds of CRCs are found in the colon and one third in the rectum.



1.2.1 The molecular basis of colorectal cancer

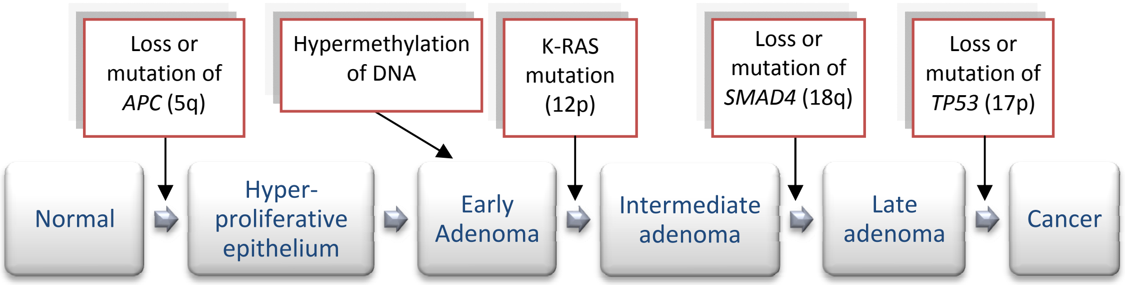
Colorectal cancers are mostly adenocarcinomas or epithelial tumours that commonly develop from polyps in the colon or rectum. The genetic pathway of most colorectal cancers is relatively well understood from the study of lesions from resected colons of Familial Adenomatous Polyposis (FAP) patients. A genetic model for the progression of colorectal neoplasms was first described by Fearon and Vogelstein in 1990 (Fearon and Vogelstein 1990) from observations of adenomas at varying stages of progression. It begins with normal epithelium and progresses through worsening levels of adenomatous dysplasia. The development of a carcinoma from normal epithelium requires the acquisition of five capabilities: the ability to replicate without external growth signals, ignore signals to stop replication, avoid apoptosis, replicate indefinitely and grow new blood vessels. Additionally a cancer can then develop invasive and/or metastatic capabilities. This development occurs through accumulated mutations in proto-oncogenes and tumour suppressor (TS) genes. Proto-oncogenes are genes that promote cell proliferation, which when mutated gain functions causing uncontrolled growth. Tumour suppressor genes are genes that normally inhibit carcinogenesis through functions such as regulating cell cycle, apoptosis, and DNA repair. Loss of function mutations cause these genes to lose this ability leading to cancer progression.

In some cases there is a strong inherited predisposition to cancer, as in FAP, where one copy of the *APC* tumour suppressor gene is inactivated in the germline. This follows Knudsons two hit hypothesis (Knudson 1971), which was formulated in the study of retinoblastoma. It describes the situation where two mutations are required for loss of TS gene function to occur, one inherited and one acquired in somatic cells later in life.

The genetic model of CRC progression generally begins with two mutations resulting in the loss of APC, hypomethylation of the DNA in epithelial cells, mutation of *KRAS* or *BRAF* oncogenes in an early adenoma (adenomatous polyp) and then loss of 18q (*SMAD4*, originally thought to be *DCC*) at intermediate (dysplastic polyp) and 17p (p53) at late adenoma (tubulovillous adenoma) stage (see Figure 1.2) (Fearon and Vogelstein 1990). It is important to note some mutations are not present in all CRC's and additional chromosomal instability and microsatellite instability can also be present (Knudson 2001).

Figure 1.2 The colorectal tumorigenesis model

The genetic model for tumorigenesis showing the changes that occur during the progression of normal epithelium to carcinoma, as proposed by Fearon and Volgelstein in 1990



Loss or mutation of *APC* (5q)
Hypermethylation of DNA
K-RAS mutation (12p)
Loss or mutation of *SMAD4* (18q)
Loss or mutation of *TP53* (17p)
Normal
Hyper-proliferative epithelium
Early Adenoma
Intermediate adenoma
Late adenoma
Cancer

1.2.2 Environmental and dietary risk factors

There is a large variation in CRC incidence between countries and it has been proposed that approximately 80% of this variability can be attributed to diet and lifestyle choices (Cummings and Bingham 1998; Parkin *et al.* 2005). The western diet is largely blamed for the increased incidence of CRC in developed countries, but there is still much to understand about the interaction between environmental and genetic factors and the effect these interactions have on complex disease risk. To this end, many studies have been conducted that focus on the effect of diet and lifestyle choices on CRC risk.

The affects of dietary risk factors in cancer have been studied using, among other prospective cohorts, the European prospective investigation into cancer and nutrition (EPIC) cohort. This is a large collection of healthy individuals aged between 45 and 74, for whom data such as diet, smoking status, alcohol consumption, and health information has been collected. Associations with increased CRC risk were reported using EPIC with low fibre intake (Bingham *et al.* 2005) and red meat (Norat *et al.* 2005), the results of this study were combined in a meta-analysis with ten similar studies that supported the association (Larsson and Wolk 2006). This was further supported recently in the EPIC Norfolk study (Park *et al.* 2010). Although this was a small case only analysis (185 cases, 62 with confirmed p53 mutation), the results showed that a higher daily intake of red meat was significantly associated with p53 mutation in tumours of late Duke's stage. This suggests an increased rate of tumour progression by p53 mutation leading to more advanced tumours. The most common somatic mutations are in the tumour suppressor gene p53 and are thought to occur at the time of progression of an adenoma to cancer (Fearon and Vogelstein 1990).

EPIC was also included in a large meta-analysis of 31 studies, which found a significant association between obesity and CRC risk (Moghaddam *et al.* 2007). Additional prospective studies have also found a link between alcohol intake and increased CRC risk (Cho *et al.* 2004).

A summary of the extensive evidence on the chemopreventive action of calcium, folate and vitamin D promoted the idea that deficiency in these compounds could increase CRC risk and highlighted plausible candidate genes and pathways for risk alleles (Lamprecht and Lipkin 2003). Vitamin D has roles in the regulation of the cell cycle and apoptosis, and folate is important in DNA biosynthesis and methylation. Intracellular calcium concentration can regulate apoptosis (Hajnoczky *et al.* 2003) and extracellular calcium has been shown to reduce β-catenin expression and increase E-cadherin expression in tumour cells from the colon (Chakrabarty *et al.* 2003). Increased β-catenin expression is a molecular marker for CRC and E-cadherin is a tumour suppressor. However, the actual effect of these compounds on CRC risk has not been confirmed.

Additionally, certain non-steroidal anti inflammatory (NSAID) drugs have been shown prevent CRC developing. However, many show toxicity themselves, such as the cox-2 inhibitor celecoxib, which although significantly reduces the recurrence of adenomatous polyps (Arber *et al.* 2006) has been prohibited owing to cardiovascular complications in treated patients. Two other drugs, Aspirin and Sulindac, have also shown effectiveness, but the risk of gastrointestinal toxicity presents challenges for long term treatment (Half and Arber 2009).

Several studies have been undertaken to attempt to quantify the effects of dietary and lifestyle risk factors and uncover the molecular basis of these suggestive associations through identification of polymorphisms that may affect how individuals respond to environmental risk factors. Some of these studies are discussed in section 1.5.2.

1.3 Genetic predisposition genes for CRC

The existence of an inherited factor affecting cancer risk has been suspected for many years, after the discovery of a number of large families that exhibited clustering of cancer cases and other phenotypic features covering several generations, which strongly suggested an inherited predisposition. These diseases were classed as Mendelian diseases as they obeyed Mendel's laws of inheritance. These conditions were characterised phenotypically by clinical observations of patients and families with the disease and by the histology of their polyps.

The predisposition genes responsible for the Mendelian CRC syndromes described below have been largely identified through the use of linkage analysis and candidate gene studies. However, these syndromes are rare in the population and only account for a small proportion of CRC cases.

1.3.1 Familial Adenomatous Polyposis (FAP)

Familial adenomatous polyposis (FAP; MIM 175100) is a dominantly inherited cancer predisposition that affects 1 in 10,000 individuals and it is characterised by the presence of hundreds to thousands of adenomatous polyps in the colon and eventual colon cancer, by an average age of 39, if left untreated. FAP patients can also develop extra-colonic manifestations such as upper gastrointestinal tumours, desmoids and congenital hypertrophy of the retinal pigment epithelium (CHRPE) (Jasperson *et al.* 2010). FAP was first described in 1925 (Lockhart-Mummery 1925). The mutation was uncovered as a deletion in the chromosomal band 5q21 in 1986 (Herrera *et al.* 1986) and the actual gene mapped to *APC* through linkage analysis of FAP families in 1987

(Bodmer *et al.* 1987). *APC* is a tumour suppressor gene that acts in the WNT signalling pathway and is mutated in approximately 80% of all tumours, which is an important step in tumorigenesis.

FAP follows Knudson's 2-hit hypothesis (Knudson 1971), where the loss of activity of APC is determined by an initial inherited germline mutation in *APC* followed by a somatic mutation (or second hit).

A variation on classical FAP is attenuated FAP (AFAP), which is caused by mutations in the *APC* gene that leave some functionality in the expressed protein. AFAP patients typically develop up to 100 adenomas in the colon. The location of the mutation is an important indicator of the severity of the phenotype (Sieber *et al.* 2006).

1.3.2 Hereditary Non-Polyposis Colon Cancer (HNPCC)

HNPCC (also known as Lynch syndrome) is a dominantly inherited predisposition to CRC with incomplete penetrance that accounts for approximately 3% of all CRC cases. HNPCC is caused by mutations in the mismatch repair (MMR) genes, such as *MSH2* (Fishel *et al.* 1993; Leach *et al.* 1993), *MLH1*(Bronner *et al.* 1994; Papadopoulos *et al.* 1994), *MSH6* (Miyaki *et al.* 1997) and *PMS2* (Nakagawa *et al.* 2004; Thompson *et al.* 2004). MMR genes play an important role in the fidelity of DNA replication through the rapid identification of mismatched nucleotides and repair of any mistakes by incorporating the correct nucleotide in the new strand (Watson *et al.* 2004). Affected individuals also have an increased risk of developing endometrial cancer, accounting for 2.3% of all endometrial cases. In fact, females with mutations in *MLH1* or *MSH2* mutations have a higher lifetime risk of developing endometrial cancer than of CRC (Resnick *et al.* 2009). The Amsterdam I criteria was established in 1991, and revised in

1999 (II), to aid the diagnosis of individuals with HNPCC and showed a specificity of 78% in the CRC population (Syngal *et al.* 2000). The Amsterdam II criteria includes: at least 3 relatives affected with an associated cancer (bowel, enodometrium, small bowel, renal pelvis or urethra) at least one of which is a first degree relative, disease affects at least two generations with one individual diagnosed under 50 years, the exclusion of FAP and pathologist verified tumours.

The Bethesda guidelines were created in 1997, and revised in 2002, to incorporate the utility of microsatellite instability (MSI) testing of tumours. The National Cancer Institute recommends MSI testing on the following microsatellite markers that are in regions of the genome not thought to be associated with cancer biology: BAT25 BAT26, DS5123, D5S346 and D18S346 (Boland *et al.* 1998). Mutations in MLH1 or MSH2 are evident in the MSI tumours of patients, but this is not always the case for MSH6.

1.3.3 The Hamartomatous Polyposis Syndromes

The hamartomatous polyposis syndromes are a group of dominantly inherited cancer predisposing conditions that exhibit similar clinical features.

Peutz-Jeghers syndrome

Peutz-Jeghers syndrome is characterised by pigmentation of the lips and fingers are predisposed to hamartomatous polyposis of the intestinal epithelium and numerous cancers including colorectal, breast and pancreatic. Affected individuals have an 81-93% chance of developing CRC (50% for breast and 11-36% for pancreatic cancer) in their lifetime (Gammon *et al.* 2009). The location of the disease gene was mapped to chromosome 19p using loss of heterozygosity (LOH) analysis on tumours and linkage analysis in 1997 (Hemminki *et al.* 1997). The gene responsible is the tumour suppressor gene serine/threonine kinase II (*STK11* (*LKB1*)) identified by the same group in 1998 (Hemminki *et al.* 1998).

Juvenile Polyposis

Juvenile polyposis (JPS; MIM 174900) does not share the physical attributes of Peutz-Jeghers syndrome, but is characterised by the presence of many juvenile polyps, a type of hamartomatous polyp, in the colon and elsewhere in the gastro-intestinal tract. Affected individuals have an increased risk of gastro-intestinal hamartomatous adenomas and cancer with a lifetime risk of 39% (Brosens *et al.* 2007). Juvenile polyposis is caused by mutations in either of the tumour suppressor genes mothers against decapentaplegic homolog 4 (*SMAD4*), which is important in transforming growth factor beta (TGF-β) signal transduction (Howe *et al.* 1998) and bone morphogenic protein receptor-1a (*BMPR1A)* (Howe *et al.* 2001). Mutations in these genes account for approximately two thirds of cases. Both were identified by linkage analysis of affected families.

Cowden Syndrome

Cowden syndrome (CS) comes under the banner of '*PTEN* hamartoma tumour' syndromes (MIM 158350). About 80% of CS cases are caused by mutations in the tumour suppressor phosphatase and tensin homolog (*PTEN*) (Marsh *et al.* 1999). The syndrome is characterised by the presence of hamartomatous polyps in the colon and multiple neoplasms of the skin, mucous membranes, breast and thyroid. CS often confused with Juvenile polyposis owing to the occasional presence of juvenile polyps in the colon of CS patients. Indeed, researchers have been known to miss-classify some

CS families as JPS, which led to the erroneous reporting of germline mutations in *PTEN* as a cause of JPS (Lynch *et al.* 1997). Members of the family in question were later found to display some phenotypic features of CS (Eng and Ji 1998). Bannayan-Ruvalcaba-Riley syndrome (MIM 153480) is another member of this group, which shares clinical features with CS and is also caused by mutations in *PTEN*.

1.3.4 MUTYH associated polyposis (MAP or MYH)

The clinical presentation of MAP (MIM 604933) shows similarity to attenuated FAP, with affected individuals developing dozens of adenomatous polyps in the colon and an increased risk of developing CRC. However, unlike the conditions described above, MAP follows an autosomal recessive inheritance model. *MYH* is a homolog of the E-coli MutY gene and is a base-excision repair gene responsible for protecting DNA in response to oxidative damage. Germline mutations in *MYH* were identified in patients with classic adenomatous polyposis with recessive inheritance. The tumours of such patients exhibit a high number of somatic G-C to T-A transversions in the *APC* gene (Sieber *et al.* 2003). The mutations in *MYH* were found through candidate gene screening of base excision repair genes after Al Tassan *et al.* described a family with an excess of G to T mutations in *APC* (Al-Tassan *et al.* 2002).

1.3.5 Hereditary Mixed Polyposis Syndrome (HMPS)

Hereditary mixed polyposis syndrome (HMPS) is a Mendelian condition that was first discovered in a family of Ashkenazi Jewish descent in 1997 (Whitelaw *et al.* 1997) and is characterised by the development of multiple colorectal polyps, of hyperplastic adenomatous or serrated adenomatous pathology, and CRC. The disease locus was found at chromosome 15q13-q14 (with maximum multipoint LOD score 4.67) through linkage analysis of a large Ashkenazi family (Jaeger *et al.* 2003). Further analysis yielded a minimal region of 10cM between markers D15S1031 and D15S118 that was shown to be highly penetrant as 18 out of 20 individuals with the haplotype were affected by the disease. This region sparked interest as it was contained within a 40cM region, known as *CRAC1*, which was detected previously in an Ashkenazi family with colorectal tumours (Tomlinson *et al.* 1999). A later comparison of these two families, plus an additional family, revealed that the 10cM minimal HMPS region was shared in all affected members of the families. This confirmed that there exists a highly penetrant CRC predisposition gene located in this region. To better define the region, eight affected individuals and one unaffected Mother of an affected offspring were genotyped using the Illumina Hap550 SNP array. The results were used to reduce the *CRAC1/HMPS* region to a minimal shared haplotype between 30,735,098 and 31,369,755 bases, which contains three known genes, the 3' end of *SCG5, GREM1* and *FMN1* (Jaeger *et al.* 2008). However, although the coding exons, promoter, introns and conserved regions were screened for variants, no mutations unique to the affected individuals were discovered.

1.4 Missing heritability of CRC

Heritability is defined as the fraction of variation that exists between individuals in a population as a result of their genotypes (Visscher *et al.* 2008). It is this variation in genotypes that renders certain individuals more at risk of developing CRC than the general population. The mutations underlying the rare Mendelian-like cancer conditions described above are highly penetrant. However, these conditions describe disease in a small number of individuals and explain just 5% of the variation in risk of CRC (Bonaiti-Pellie 1999), which leaves a large proportion of heritability unexplained.

In 2000, Lichtenstein *et al.* published the results of the analysis of 44,788 pairs of mono- and di-zygotic twins from the Netherlands (Lichtenstein *et al.* 2000). The aim of the study was to determine the relative effects of environmental and genetic factors on the heritability of cancer. This was possible with this type of study, in preference to family studies, because both types of twins share an environment from conception with mono-zygotic twins sharing 100% of genes and di-zygotic twins sharing 50% of genes. Essentially, if the results showed that mono-zygotic twins both developed cancer more often than di-zygotic twins then genetic similarities between the twins were important. However, if di-zygotic twins and mono-zygotic twins both developed cancer with similar rates then the shared environmental factors were probably important. The results showed that, for CRC, heritable factors accounted for approximately 35% (95% CI: 10%-48%) of the variance in risk, while shared environmental factors between twins accounted for 5% and non-shared environmental factors, 60%. Lichtenstein *et al.* concluded that the major contributor to cancer in this study was the environment, but that there were “major gaps in our understanding of the heritability of colorectal, breast and prostate cancer”. The study found that the mono-zygotic twin of a person with CRC had an 11% increased risk of developing the same disease by age 75, whereas this figure dropped to 5% for dizygotic twins (or siblings)(Lichtenstein *et al.* 2000). Therefore, although on a population level the increased risk caused by heritable factors was moderate; the information could be valuable in a clinical setting for relatives of individuals with cancer.

There were some limitations to this study, including the assumption that there are no interactions between genetic and environmental factors. Also, although 10,803 cancers were included in the study, the sample size for monozygotic and dizygotic concordant twins with CRC was quite small at 30 and 32, respectively, compared to discordant twins where there were 416 monozygotic and 846 dizygotic twins. This may have affected the results and the author's conclusions that most of the variance in risk for CRC was caused by environmental and non-shared risk factors rather than being inherited. The 95% confidence interval for the estimated heritability was very large at between 10% and 48% highlighting the uncertainty of the 35% estimate. Nevertheless, these findings provided the impetus for renewed efforts to find the missing heritability for CRC, and other cancers.

1.5 Methods for the detection of predisposition genes

The technique used to detect predisposition genes or susceptibility alleles depends on the model of inheritance. Before GWA studies were feasible, linkage analysis, association analysis of candidate genes, and direct sequencing were the methods of choice for the detection of genetic predisposition. Generally, candidate genes would be identified through linkage analysis and then followed up with an association analysis and the sequencing of affected individuals to determine the causal mutation.

1.5.1 Linkage Analysis

Linkage analysis was the primary method for detecting high risk genes responsible for disease in families with a large number of affected members. The method scans the genome searching for regions that are shared amongst affected individuals within a family at a level that is higher than expected. These regions segregate with the disease and can be used to identify the location of the disease gene within the boundaries of a linkage peak.

Mapping the location of genes in this way relies on the process of crossing over that occurs in meiosis during the formation of gametes. This is the process whereby homologous chromosomes form pairs, each pair being made up of four chromatids. The pair of homologous chromosomes is then separated, while one chromatid from each chromosome maintains contact at certain positions called chiasmata. These reveal the location where crossing over has occurred and two chromatids, on different chromosomes, have exchanged segments of DNA producing two recombinant chromosomes. Recombination has a semi-random nature and was crucial in the construction of genetic maps of the genome. If two loci or markers that lie on the same chromosome segregate independently then one or more recombination events must have occurred between them. This event becomes more likely the further apart the markers are located. Markers that are close together are unlikely to be separated by recombination and are normally inherited together, with other surrounding alleles, forming a haplotype block, which can be traced through pedigrees and used to locate a disease gene. The recombination fraction is the proportion of gametes that are recombinant between two loci. Two markers that segregate independently will have a recombination fraction (θ) of 0.5. In this way, θ provides a measure of the genetic distance between markers (Ott 1999).

The locations of recombination can be determined using dense panels of genetic markers. As closely related individuals share a large proportion of the genome, the number of markers required to detect linkage is relatively small. However, as the number of recombination events in a given family is low, the boundaries of a linkage signal are likely to be several megabases apart and further fine-mapping of the region is required to establish the disease gene.

1.5.1.1 Physical and genetic maps

Physical maps of the genome show the position of marker, measured by the distance from the telomere, which has been determined by physical methods. Whereas, the genetic position, measured in centimorgans (cM), should give the same order, but indicates the probability that markers will be separated by recombination. As the rate of recombination varies, the distances of the two maps will be quite different. Equally, individuals may have different genetic distances depending on the number of crossovers per meiosis.

The development of genetic linkage maps of variation in the human genome began with Botstein *et al* in 1980 with the proposed use of restriction fragment length polymorphisms (RFLPs) as markers for linkage (Botstein *et al.* 1980). Although, RFLPs are not greatly polymorphic and the technique was expensive and time consuming several Mendelian diseases were mapped in this way, including Huntington's disease (Gusella *et al.* 1983). This developed into the use of microsatellites (short tandem repeats) as genetic markers, which were much more polymorphic and abundant through the genome. This technique was also cheaper as the results could be assayed using PCR and electrophoretic separation and microsatellites became the markers traditionally used for linkage analysis. Most of the known genetic predispositions to CRC described above were detected by this method.

However, with the discovery of millions of single nucleotide polymorphisms (SNPs) throughout the genome whole genome SNP linkage techniques were developed. SNPs are less polymorphic than microsatellites, but are far more abundant and provide greater coverage across the genome. Equally, as most SNPs are bi-allelic, genotyping requires a much simpler assay and are more scalable in terms of genotyping large numbers of samples (Kruglyak 1997). Although the use of SNPs had been considered for the high density panels that would be required for association studies, they were not utilised for linkage analysis until a SNP linkage map and 3.9cM SNP set of 1,891 SNPs in 719 clusters and 332 singleton SNPs were produced in 2003 (Matise *et al.* 2003). The SNP linkage map showed a high level of concordance with the deCode and Marshfield microsatellite (or STR) linkage maps that were already available and provided the tool required to bring SNPs to genome screening by linkage (Broman *et al.* 1998; Kong *et al.* 2002).

1.5.1.2 When is a signal considered significant?

In linkage analysis, the parameter of interest is θ between two markers, i.e. the probability of recombination between two loci during meiosis. The null hypothesis is that there is no linkage between a marker and disease locus (θ will equal 0.5). The alternative hypothesis, the presence of linkage, is indicated by a θ value less than 0.5. There are two main ways to conduct a linkage analysis, two-point and multipoint linkage. Two-point linkage mapping calculates the recombination fraction between a marker and the disease. Multipoint linkage is more efficient as it examines multiple markers at the same time and determines the location of a disease locus in relation to a map of markers.

The odds of linkage are calculated by the ratio of the likelihood of the pedigree if the loci are linked against the likelihood of no linkage. The logarithm of the odds provides the LOD score, of which the most likely recombination fraction gives the maximum LOD score. The results of a linkage scan are, thus, a list of LOD scores defined by the locations of markers. Generally, a LOD score will be considered significant if it is greater than 3.3 (Lander and Schork 1994). Linkage at the locus can be rejected if the LOD score is lower than minus two.

In the presence of linkage with the disease phenotype, the results will identify a relatively large region of the genome. The size of the region depends on the location and number of crossovers between the marker and the disease. Recombination events result in the shared region reducing in size through the generations allowing for better mapping of the actual disease locus. In the absence of many large multi-generation families the number of crossovers will be small, resulting in a large region of linkage (Boehnke 1994). In order to increase the number of crossovers, and more finely map the disease locus, the number of generations would need to be very large.

Linkage is powerful for the detection of highly penetrant dominantly inherited disease alleles in related individuals. However, families need to be big to provide enough power to detect a signal and the technique loses power if the effect size falls below two.

1.5.1.3 Efforts to detect further CRC predisposition genes by linkage analysis

Further autosomal dominant predisposition genes for CRC have been sought using this method. In 2003 a significant region of genetic linkage was detected on chromosome 9q22.2-31.2 using 53 families where at least two siblings were affected with the disease (Wiesner *et al.* 2003). The results of this study were supported in an independent study of 57 CRC families recruited as part of the colorectal tumour gene identification (CORGI) study from the UK, which refined the region to 9q22.32-31.1 (Kemp *et al.* 2006). A genome-wide linkage study of 69 CRC families from the CORGI study identified an additional locus on chromosome 18q21 was also identified with a maximum non-parametric LOD score of 3.1. Restricting the affection status criteria to only include those with CRC diagnoses identified another locus on chromosome 3q21-q24 (maximum non-parametric LOD score of 3.4) (Kemp *et al.* 2006). These results lend weight to the existence of additional CRC predisposition genes. However, despite screening all genes in the 3q21-q24 locus no potentially causal variant was identified to explain the linkage signal. The genes responsible for these linkage peaks have yet to be identified.

1.5.2 Direct Association studies by candidate genes and SNPs

Initial direct association studies were, by necessity, performed on candidate regions that contained genes likely to be involved in the disease owing to their function. However, candidate gene analysis methods are still a complementary technique in fine-mapping loci identified by indirect association studies. The markers most commonly used are single nucleotide polymorphisms (SNPs), which are well characterised, abundant in the genome and less mutable than microsatellites.

Selection of candidate genes is difficult and the approach is limited to genes of known function, which ignores a large proportion of the genome. The SNPs that were typed were generally non-synonymous and/or in regions of potential functional importance and resulted in changes likely to have a direct biological effect. The polymorphisms were identified through the study of candidate genes that fall within pathways involved in cancer progression.

In 2001, Houlston and Tomlinson reviewed the reported associations between polymorphic variation and colorectal cancer risk and performed a meta-analysis of the published studies (Houlston and Tomlinson 2001). Although some of these associations are probably real, the meta-analysis showed mixed results and suffered from individually small sample sizes and mixed ethnicity leading to difficulty replicating the original findings. Another issue with meta-analyses of this sort is publication bias in the sense that negative results may not have been published, while small studies showing positive results have. This is a possibility for some of the variants published in this paper, including the MTHFR and APC I1307K polymorphisms that are discussed below.

In a more recent review of published candidate gene association studies on the genetic susceptibility to cancer, 344 reported gene variant associations from 161 articles on a number of different cancers were investigated (Dong *et al.* 2008). The authors analysed the reports using a false positive report probability (FPRP) method, which uses the probability that a finding is false given a statistically significant result to give a measure of how likely the result is for a given study. The results of this study left thirteen significant associations at a prior probability of 0.001 that were mostly related to metabolising enzymes. Four gene associations remained significant at a level similar to that used for association studies (10^{-7}) and are much less likely to be false positives than the other studies. These were MTHFR C677T in gastric cancer, NAT2 slow acetylator phenotype in bladder cancer, and GSTM1 null in bladder cancer and Leukaemia. These genes overlap with reported associations relating to CRC and these have been separated by function and discussed further in the sections that follow.

1.5.2.1 Carcinogen metabolising enzymes

The increase in risk of CRC associated with excessive red meat intake is based around the variation in activity of carcinogen metabolizing enzymes. Two enzymes important in the metabolism of aromatic and heterocylic amines, such as those found in cooked meat, are N-acetyltransferase 1 (NAT1) and 2 (NAT2). There is a high frequency of polymorphic variation in the genes encoding these enzymes and the alleles of certain polymorphisms translate to a rapid or slow acetylator phenotype. The rapid acetylator phenotype is associated with an increased risk of CRC (Roberts-Thomson *et al.* 1996). However, there are difficulties quantifying an individual's exposure as this is dependent on the accuracy of the answers to questionnaires. More recent studies by Hein *et al* showed that certain polymorphisms reduced enzyme activity through reduced expression leading to a slow acetylator phenotype in *S. Pombe*, but also that activity may be regulated by the concentration of substrate (Hein 2002).

1.5.2.2 Oncogenes and tumour suppressors

The attenuated form of FAP is caused by mutations in certain locations of the *APC* tumour suppressor gene and other polymorphisms were predicted to possibly increase the risk of CRC through a similar route (Bonaiti-Pellie 1999). Two such variants have been reported to increase CRC risk, I1307K, which was found in the Ashkenazi Jewish population (Laken *et al.* 1997), and E1317Q (Frayling *et al.* 1998; Lamlum *et al.* 1999).

1.5.2.3 Methylation Genes

Folate metabolism has been suggested to affect cancer risk by influencing the availability of methyl groups, which could in turn affect DNA methylation and hence the expression of proto-oncogenes and tumour suppressors. Polymorphisms in the enzymes 5,10-methylenetetrahydrofolate reductase (MTHFR) and methionine synthetase (MTR) that affect folate metabolism are good candidates for CRC risk alleles and have been comprehensively studied with mixed results.

In a recent meta-analysis of the published studies on the MTHFR C677T polymorphism, Taioli and colleagues reported that out of 29 studies just two showed a significant inverse association of the TT genotype with CRC risk that did not have odds ratios (OR) with 95% confidence intervals spanning the value 1 (Taioli *et al.* 2009). The meta-analysis of 13,992 samples over fourteen studies gave an overall OR of 0.83 (95% CI: 0.74-0.94) and a P value of 0.003. However, Eussen *et al.* recently described a case control study on plasma folate levels, polymorphisms and CRC risk in 1,367 cases and 2,325 controls, matched by age, gender and geographic centre, from the EPIC cohort (Eussen *et al.* 2010). The study found no association between CRC and any of a number of MTR and MTHFR polymorphisms, including MTHFR C677T. Clearly, the jury is still out on whether the effect of this polymorphism significantly increases CRC risk.

1.5.3 Candidate Gene Screening

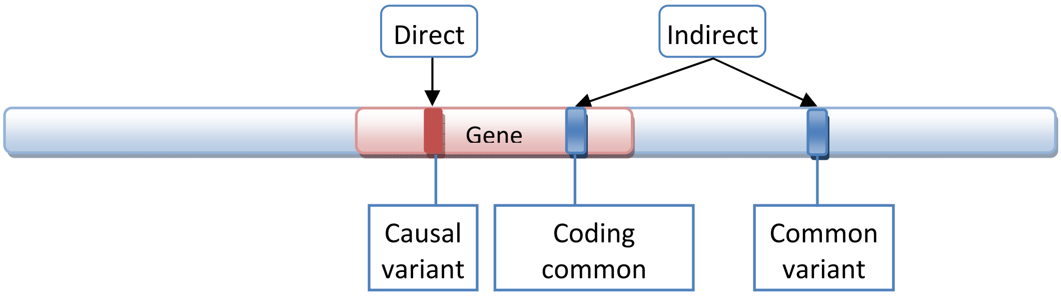
Once a candidate gene has been identified, by linkage or function, the only real method available to detect the actual causal variant is to sequence the gene in cases and controls. Until very recently, this has been somewhat limited to exons and coding regions of candidate genes owing to the time consuming nature of standard sequencing. However, the development of inexpensive and rapid next-generation sequencing technologies is allowing whole regions and even entire genomes of individuals to be sequenced (Metzker 2010). This technique will be valuable in the search for rare variants that influence complex disease phenotypes and has been utilised in the 1000 genomes project, which aims to sequence 1000 genomes to detect rare and undiscovered variants with a frequency greater than 1% in the human genome (www.1000genomes.org). This project aims to provide an additional reference from which to predict the untyped genotypes of millions of variants, discussed later, and also to allow better fine-mapping of regions detected by association studies.

1.6 Indirect association studies and linkage disequilibrium

The methods discussed above largely involve direct analyses where the causal variant has been genotyped and analysed directly. However, genome-wide methods generally rely on analysing a subset of the variants in a region that can be used as proxies for the total variation. These studies are known as indirect methods (see Figure 1.3)

Figure 1.3 The indirect and direct methods of causal variant detection

Some of the possible positions of the genotyped common variants in relation to the causal variant are indicated. In a direct association analysis the causal variant is genotyped and analysed directly. However, in an indirect method, the genotyped variants are in LD with the causal variant and so the analysis identifies its approximate location. Therefore, the actual genotyped variant could be coding or non-coding and synonymous or non-synonymous. The causal variant is shown within a gene, but could be located in a regulatory region that regulates a gene nearby or a gene on a different chromosome.



Common SNPs are the result of historical mutations in a population and are associated with other alleles that happen to be in the same region of the same chromosome. These alleles form a haplotype and will be inherited together in subsequent generations, unless split by recombination events or crossing over between parental chromosomes. The alleles of the haplotype are in linkage disequilibrium. That is they are "found together on the same chromosome more often than expected if they were segregating independently" (Ardlie *et al.* 2002). The first suggestions that LD could be used to find disease association with a marker were from studies of HLA and Hodgkin's disease (Bodmer 1973). This paper led to the suggestions that linkage disequilibrium could account for association between disease and a genetic marker by LD. Mutations were discovered by identifying a locus associated with the disease and then identifying the causal mutation.

Interest in the elucidation of linkage disequilibrium was reignited by unsuccessful linkage studies for complex diseases and the need for dense genome-wide SNP based association studies. The identification of haplotype blocks of long range LD structure (Daly *et al.* 2001) showed that it is not necessary to type all markers in all genes to find an association with disease as LD relationships between SNPs allow the genotyping of a select subset of SNPs that can be used to effectively tag all known variation, in a method of genome wide indirect association. The use of this method leads to an unbiased search for susceptibility variants as no prior knowledge of gene function or suitability as a candidate for the disease is required. This can lead to the discovery of new genes and molecular pathways that were not suspected of having any involvement in the disease.

The two chief measures of LD of interest to association studies are both based on Lewontin's statistic D, which measures the difference between observed frequency and expected frequency under random segregation (Lewontin 1964).

$$D = P_{AB} - P_A \times P_B$$

Where P_{AB} is the frequency of the AB haplotype and P_A and P_B are the allele frequencies at each of the two alleles. D is highly dependent on allele frequency and so D' and r^2 are used as alternative measures of LD. D' is calculated by dividing D by the maximum possible value given the allele frequencies. If D'=1, then the markers are in *complete* LD and there has been no recombination between them. However, if D' is less than 1, this implies that LD is not complete, but the actual level of LD is difficult to interpret and can be heavily influenced by the number of samples, where small sample sizes lead to inflated estimates.

The usual method for determining suitable tagging SNPs for association studies is r^2 or the correlation coefficient. If $r^2=1$ between markers, then they are in *perfect* LD and have the same allele frequency. Therefore, one marker could be used as a proxy for the other. Values of r^2 below one do not necessarily indicate recombination between markers, as differences in allele frequency will have the same effect. The value of r^2 provides a measure of the information that one marker will provide on another and the power of that marker to detect an association with disease caused by another functional variant in LD with it. The sample size would need to be increased by $1/r^2$ to achieve the same power to detect an association with a SNP in LD with the causal variant as would be achieved by typing the causal variant directly (Ardlie *et al.* 2002; Morris and Cardon 2007).

LD structure varies between populations and can be influenced by population structure, admixture (or migration of individuals from different populations that might have different allele frequencies), genetic drift (changes in haplotype frequency caused by random sampling of available gametes in each generation) and variable mutation and recombination rates between populations (Ardlie *et al.* 2002). Therefore, knowledge of LD structure in different populations is vital for the design of panels of SNPs to provide whole genome coverage in GWA studies.

1.7 The development of GWA for complex disease

1.7.1 A move away from linkage analysis

Linkage analysis is a powerful method for the detection of rare, high to moderate penetrance mutations and was very successful for Mendelian conditions or dominantly inherited conditions. Genome-wide linkage SNP panels have been applied to complex diseases and identified variants associated with disease, for example the variants identified in inflammatory bowel disease and Crohn's disease (Hugot *et al.* 2001; Ogura *et al.* 2001; Rioux *et al.* 2001; Stoll *et al.* 2004). However, the method has had limited success, as the detection of common variants with a small effect size requires large multi-case families and the recruitment of multiple generations, which is not always possible in late-onset complex diseases. More importantly, once penetrance falls below about 10%, as in the common cancers, linkage analyses lack statistical power.

Penetrance is the probability that an individual will develop a certain phenotype, or disease, if they carry a certain genotype. In Mendelian diseases it is usually clear when you have found the causative mutation, as it is absent in controls and unambiguous, e.g. affects protein function. However, low-penetrance or low risk mutations present problems for linkage as they do not fully segregate with disease; there will be unaffected individuals that carry the mutation and affected individuals that do not. This is called incomplete penetrance and leads to a diminished linkage peak. Therefore an alternative strategy is required for these low risk variants.

Additional reasons leading to a lack of power in linkage are heterogeneity, where the phenotype is caused by different genes in different families and cancels out any linkage when families are combined, and the presence of phenocopies or individuals that present the phenotype, but lack the associated genotype that is shared by other affected family members (Pharoah *et al.* 2004).

In 1996, Risch and Merikangas (Risch and Merikangas 1996) noted that, despite a large number of reports on the genetic basis of complex diseases, only a small number were replicated in subsequent studies. The authors determined that linkage analysis, while successful for highly penetrant dominantly inherited conditions, was underpowered to detect the smaller effects of the genes responsible for complex diseases. They calculated that once the relative risk of a locus dropped below 2, the number of families required to detect it was greater than was feasible to obtain. Association studies on candidate genes and polymorphisms, on the other hand, had much greater power and were shown as much better suited to the search for genes with smaller effect sizes thought to underlie complex diseases (GRR of 1.5 required a sample size of less than 1000). It was noted that several advances would be required to move away from the candidate gene to a genome wide association study, principally the completion of the human genome sequence, identification of polymorphisms across the genome, and the ability to genotype large numbers of samples for hundreds of thousands of polymorphisms.

1.7.2 The human genome project, HapMap and mapping genetic variation

With the completion of the human genome projects in 2001 (Lander *et al.* 2001; Venter *et al.* 2001) came the publication of the draft human genome sequences that would allow studies into genetic variation that exists between individuals, genetic architecture and patterns of linkage disequilibrium (LD) across the genome. Ninety per cent of the variation that exists between individuals is caused by single nucleotide polymorphisms (SNPs), where the DNA sequence at one base pair varies and can have two alternative alleles (Collins *et al.* 1998). Such alterations, especially within the coding regions of genes or regulatory sequences can alter the function of a gene resulting in an increased susceptibility to certain common diseases or changes to drug metabolism. Most SNPs do not occur in these regions, but provide good markers for variation owing to regions of LD.

In 2001, the International SNP working group published a map of 1.42 million SNPs that included sequence information and the genetic and physical positions of the SNP in the genome (Sachidanandam *et al.* 2001). These SNPs were a union of The SNP Consortium's 1,023,050 SNPs studied in publically available data from 24 ethnically diverse samples from the DNA Polymorphism Discovery Resource (Collins *et al.* 1998; Altshuler *et al.* 2000) and those from the International Human Genome Sequencing Consortium (Mullikin *et al.* 2000). This was the first available genome-wide map of all known SNPs that could be used to explore haplotype diversity among populations and disease susceptibility. It is important to note that these SNPs originated from the differences between the two genomes sequenced in the Human Genome Project and more data would be needed to more completely identify the true level of polymorphic variation in humans.

Estimations suggested that there were around 10 million loci with variation at a frequency greater than 1% that make up 90% of observed variation with the remainder consisting of rare variants (Kruglyak and Nickerson 2001; Reich *et al.* 2003). In order to move forward with genome-wide association, more information was needed on the allele frequency of variants in different populations and patterns of LD to enable the selection of appropriate markers to tag the genome.

The HapMap project progresses this initiative. The aim of the HapMap project was to "determine common patterns of DNA sequence variation in the human genome, by characterising sequence variants, their frequencies and correlations between them, in DNA samples from populations with ancestry from Africa, Asia and Europe" (The International HapMap Consortium 2003). The data from this project formed the necessary tools to perform genome-wide association studies and tests of indirect association, where prior evidence for the probable involvement of a variant is not required. Owing to technical constraints, association studies were limited to candidate genes and variants of known function. The HapMap provided the necessary information to design SNP panels to cover the genome owing to knowledge in LD and the information on differing allele frequency and LD relationships between different populations allowed for the better design of association studies and to overcome the problems of population stratification.

The HapMap consortium reported phase one of the project in 2005 (The International HapMap Consortium 2005) where 1.3 million SNPs were genotyped in three populations. The samples consisted of 90 CEU containing 30 parent-offspring trios (with Northern or Western European ancestry, from Utah, USA and part of the Centre d'Etude du Polymorphisme Humain Collection (CEPH)), 90 YRI (from the Yoruba, Nigeria), 45 CHB (from Beijing, China) and 44 JPT (from Tokyo, Japan). The HapMap was later extended to 3.1 million SNPs in phase II (The International HapMap Consortium *et al.* 2007) and has since been expanded with additional samples from different populations.

1.7.3 High density tagging SNP genotyping chips

In 2005, both Illumina and Affymetrix published papers describing their methods for high density genome wide SNP genotyping arrays. Affymetrix produced the Human Mapping 500K array (Di *et al.* 2005). The SNPs were randomly chosen, but evenly spaced across the genome (approximately 2.5kb apart). The Affymetrix product datasheet shows that this chip had 70% coverage of the HapMap2 SNPs at $r^2>0.8$.

Illumina developed a different method for their high throughput genotyping array (Gunderson *et al.* 2005). The method involves bead arrays that contain probes for each SNP (one for allele A and one for allele B). When DNA is hybridised to the beads it undergoes allele-specific primer extension and is labelled with biotin, which is then detected through immunohistochemistry to read the genotype. The Illumina Hap300, 550 and 1M SNP genotyping arrays consist of SNPs chosen from the HapMap to best tag the entire genome for SNPs that have a minor allele frequency greater than 0.05. The product technical note: "the power of intelligent SNP design", shows that the Hap550 SNP array provides 90% coverage of HapMap CEU loci with $r^2>0.8$.

The production of highly parallel genotyping methods reduced the cost of genotyping thousands of samples for hundreds of thousands of SNP dramatically and made high-density GWA studies possible.

1.8 The Genome Wide Association Study

Genome wide association (GWA) studies are indirect methods of association often performed using tagging SNPs and attempt to map the location of disease genes, or susceptibility alleles, through the comparison of allele frequencies of markers between individuals with the disease (cases) and healthy individuals without (controls). The GWA study is well suited to the detection of low penetrance susceptibility alleles that confer a moderate risk of disease. The markers most commonly used are dense panels of common SNPs, ideally chosen to reflect the LD structure in the population of study, to give adequate genome coverage. The development of technology and reduction in cost made it feasible to genotype thousands of samples over 550 thousand to 1 million genome-wide SNPs.

1.8.1 The assumptions and limitations of GWA Studies

In order for GWA studies to be successful a major assumption must be met. The power to detect an association using this method will be greatly reduced if there is allelic heterogeneity at the disease locus between affected individuals, i.e. the disease is caused by varying mutations in the same gene in different individuals (Pritchard 2001). There has been some surprise that the loci identified in most published GWA studies are non-functional synonymous SNPs and that most do not map to known genes (Hardy and Singleton 2009). However, given that these are indirect association studies and the SNPs genotyped, chosen based on their ability to tag the genome, are a fraction of the total number of common variants, this should not be unexpected. As only a small proportion of SNPs are synonymous and located in regulatory or coding regions of genes, it is unlikely that causal SNPs will be genotyped directly. Instead, the top associated SNPs are likely to be one or more markers that are in high LD with the causal variant. The causal variant could be an untyped common SNP, but could also be a structural variant or rare variant.

The goal of the GWA study is to determine associated susceptibility alleles and identify regions associated with disease. These regions provide candidate genes for further analysis and sequencing to find functional changes that are in LD with the identified SNPs. Although the causal variant should be in high LD with the top genotyped SNP, it does not necessarily follow that the affected gene is in the same LD 'block' or even in the vicinity of the casual variant. The causal variant could be located within the coding region of a gene or regulatory element in the same LD 'block' as the typed SNP, but could also affect a gene some distance away, but regulated within the same pathway (Ioannidis *et al.* 2009). Therefore, finding the causal allele will require further fine mapping of the regions detected, information on gene pathways and the use of sequencing technologies to uncover.

1.8.2 Overall GWA study hypothesis

GWA studies work under the assumption of the 'common disease common variant' (CDCV) model that proposes that a considerable proportion of the variance in risk for complex disease is caused by common variants with modest effect sizes (Cazier and Tomlinson 2009). Thus, to determine the missing heritability of complex diseases GWA studies were designed to have power to detect common disease variants (MAF greater the 5%) with effect sizes between 1.3 and 1.5. The results from GWA studies to date have shown this hypothesis to be a valid one, with the publication of over 500 significant common SNP associations (National Human Genome Research Institute Catalogue of published GWA studies at http://www.genome.gov/26525384).

The main alternative theory to this is the ‘common disease rare variant’ (CDRV) hypothesis, which suggests that a large proportion of the variance in risk of common diseases is caused by the combined effects of a number of rare variants with moderate effect sizes (Bodmer and Tomlinson 2010). However, until very recently, the detection of rare variants has been hindered by the speed and cost of the available technology. As susceptibility variants have low penetrances, they rarely show familial clustering and will, therefore, be difficult to detect using familial studies. Using the GWA study design, one would require more than 70,000 cases and controls to have 80% power to detect a variant with a frequency of 0.05% and a relative risk of 1.5 (Carvajal-Carmona 2010). Very few GWA studies would have sufficient power to detect variants with a MAF of less than 0.05 and alternative strategies are required for their detection, such as sequencing large number of cases and controls.

1.8.3 Population stratification

Population stratification is the presence of subgroups of samples within the data with differing SNP allele frequency and disease incidence owing to the inclusion of different ethnic populations. This is especially a problem if the cases are from a different population than the controls leading to significant differences between them that are not caused by disease, but are false positives (Hirschhorn and Daly 2005). Association studies have had bad press in the past owing to poor replication of significant results, which had largely been attributed to population stratification or structure. However, the problem was largely exaggerated as few studies have been published where this has been a major issue (Cardon and Palmer 2003) and the lack of replication is likely due to poor study design, small sample sizes and publication bias.

However, as sample sizes have increased, it has become clear that matching ancestry within datasets is vital to avoid false positive associations (type one error). Evidence for the effect that differences in allele frequency have on association results was demonstrated in a study on height in European Americans. The study found a significant association between height and the functional SNP in the *lactase* gene (*LCT)* that is responsible for lactase persistence. The allele frequency of this SNP varies from 0.2 to 0.8 across European populations (Campbell *et al.* 2005). The study highlighted the presence of stratification caused by this large variation in allele frequency, which was not detected by standard methods of genomic control or the program STRUCTURE. The ability to detect stratification depends on the SNPs and the number of samples used. The increase in knowledge of allele frequencies in different populations and LD structure has led to advanced methods for the detection and correction of stratification in GWA studies by principal components analysis (Price *et al.* 2006).

1.8.4 When is an association signal significant?

The accepted genome wide significance level for association is 5×10^{-8} (Risch and Merikangas 1996), which makes use of the Bonferroni correction for the number of SNPs tested (1 million independent SNPs). Although this reduces the number of false positives (a P value of 0.05 would produce significant results for 5% of SNPs by chance which is 27,500 false positive associations in 550,000 SNPs), it loses power. The method is a stringent one as it assumes that all tests are independent, which is not true for high density SNP panels owing to LD between SNPs (Hirschhorn and Daly 2005).

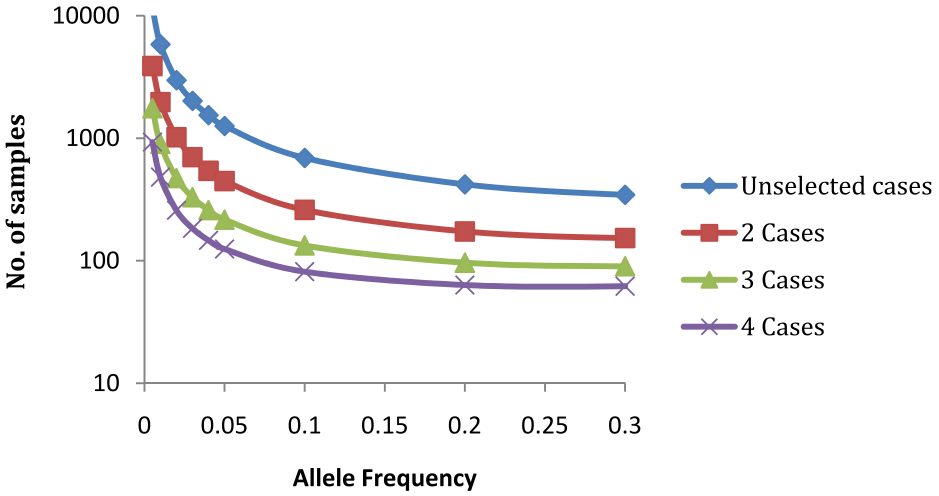
The replication of significant results in an independent dataset is vital to ensure that the finding is real and not just an artefact of the initial dataset. This is often achieved by genotyping a large number of SNPs (550K) in a modestly sized first stage, to determine a subset of SNPs showing evidence of association. These SNPs are then genotyped in a much larger second phase, or replication phase, using a more stringent Bonferroni corrected significance value. The strategy is highly efficient in minimising the amount of genotyping required and hence reducing the cost. This allows more samples to be genotyped in the second phase and maintains power (Satagopan *et al.* 2002). To increase power further, the two stages can be analysed jointly by meta-analysis and a third dataset added for a replication phase (Skol *et al.* 2006).

We have undertaken a large multi-stage GWA study with the aim of identifying common alleles that lead to an increase in CRC risk. To maximise the power to detect an association, each case in phase one had at least one first-degree relative with CRC. Cases with a family history of the disease greatly increase the chance that an affected individual will carry the susceptibility allele. Compared to association studies based on cases not selected for family history, the number of cases required to detect a risk allele reduces by about two-fold if cases have one affected first degree relative and more than four-fold with two (see Figure 1.4). A study consisting of 1000 cases and

1000 controls will have 90% power to detect variants with a relative risk of two and an allele frequency greater than 0.05, if cases have two affected relatives. This allowed us to genotype a smaller enriched dataset for the maximum number of SNPs in the first phase and a much larger dataset for the most associated SNPs in phase two, thus, providing a more efficient and cost effective approach. The design of our GWA study is discussed more fully in Chapter 3.

Figure 1.4 The number of samples required with varying numbers of affected relatives

The graph shows the number of samples to detect an allele with a relative risk of 2, with varying numbers of affected relatives (with matching cases and controls, power=90% and α=0.001).



1.8.5 Overview of the recent successes from GWA studies in CRC

Later in this thesis, I describe the study design and analysis of the CRC GWA study conducted in our laboratory and the research that has followed in the last four years. However, an overview of the recent GWA study results for CRC, including those from other groups, is given below (Table 1.1). All of the associated SNPs have been confirmed in additional studies and most are located in regions near genes with plausible functional effects or genes that act in pathways that could affect CRC risk.

The first reported associated SNP from a GWA study in CRC was rs6983267 at chromosome 8q24.21. This SNP was also discovered in independent studies to be significantly associated with prostate cancer (Haiman *et al.* 2007) and ovarian cancer (Ghoussaini *et al.* 2008) suggesting that this variant may influences general cancer risk.

Table 1.1 The reported SNP associations from the CRC GWA Studies

SNP	Chr.	Position (bp)	Combined P value	OR	Reference	Gene region
rs6983267	8q24.21	128,482,487	$1.27\text{x}10^{-14}$	1.21	(Tomlinson *et al.* 2007)	POU5F1P1
rs10505477	8q24.21	128,476,625	$3.73\text{x}10^{-14}$	1.12	(Tomlinson *et al.* 2007) (Zanke *et al.* 2007)	DQ515898
rs7014346	8q24.21	128,493,724	$8.6\text{x}10^{-26}$	1.19	(Tenesa *et al.* 2008)	DQ515898
rs4939827	18q21.1	44,707,461	$1.00\text{x}10^{-12}$	0.85	(Broderick *et al.* 2007)	SMAD7
rs4779584	15q13.3	30,782,048	$4.7\text{x}10^{-7}$	1.23	(Jaeger *et al.* 2008)	-
rs16892766	8q23.3	117,699,995	$3.3\text{x}10^{-18}$	1.25	(Tomlinson *et al.* 2008)	EIF3H
rs10795668	10p14	8,741,225	$2.5\text{x}10^{-13}$	0.89	(Tomlinson *et al.* 2008)	-
rs3802842	11q23.1	110,676,919	$1.08\text{x}10^{-12}$	1.17	(Pittman *et al.* 2008) (Tenesa *et al.* 2008)	C11orf53
rs9929218	16q22.1	67,378,447	$1.2\text{x}10^{-08}$	0.91	(Houlston *et al.* 2008)	CDH1
rs4444235	14q22.2	53,480,669	$8.1\text{x}10^{-10}$	1.11	(Houlston *et al.* 2008)	BMP4
rs10411210	19q13.1	38,224,140	$4.6\text{x}10^{-09}$	0.87	(Houlston *et al.* 2008)	RHPN2
rs961253	20q12.3	6,352,281	$2.0\text{x}10^{-10}$	1.12	(Houlston *et al.* 2008)	BMP2

The current results show that the CDCV hypothesis is true, but also that the situation more complicated than that. These variants account for about 6% of the variance in risk and additional variants remain to be discovered. All of the GWA study identified SNPs confer a low relative risks of disease and it is likely that the total variance is not fully covered by one hypothesis and that rare variants, and structural variants, will also contribute to risk, although their detection and analysis is more difficult.

1.9 Applications beyond basic association

The genotyping of thousands of genome wide tagging SNPs for GWA studies has generated a large amount of genetic data from cases and controls that can be used for more than just the basic association tests. In addition to the imputation of untyped SNPs, discussed below, there are numerous population genetic methods that can be applied to the data to explore alternative routes in determining or fine mapping disease association. These include haplotype analysis, homozygosity mapping, detection of population structure and the detection of structural variation such as copy number duplications and deletions.

1.9.1 Predicting genotypes at untyped SNPs

As the need has arisen for increasing numbers of samples to detect variants of modest effect, GWA studies have increased in size through numerous collaborations with other groups that have collected similar datasets of cases and controls. Owing to the number of different SNP arrays available, the overlap in genotyped SNPs can vary greatly between studies. Imputation delivers a solution to generate improved SNP overlap between studies through the prediction of missing genotypes based on a reference panel, such as the HapMap. The resultant genotype probabilities can then be combined in a meta-analysis to improve the power to detect susceptibility variants. An additional use of imputation is to help fine-map the region of an association signal to identify additional associated SNPs that have not been genotyped. There are a number of different programs to undertake imputation, including MACH, BEAGLE, PLINK and IMPUTE and numerous comparisons between the methods have been reported in the literature. Overall, MACH and IMPUTE seem to achieve similar results, although IMPUTE has a higher accuracy, and both exceed the alternative options (Pei *et al.* 2008).

The ability of these programs to accurately predict genotypes at untyped SNPs depends on the SNPs that have been directly assayed. Imputation will only be successful if the genotyped SNPs are in high LD with the SNPs being predicted. In this way, genotyping chips that have been designed to include SNPs that best tag the genome based on LD with other common variants (such as the Illumina 550 or 300K) outperform those that were designed with SNPs that were evenly spaced across the genome (such as the Affymetrix 500K chip). Another aspect that affects the accuracy of imputation is minor allele frequency of the SNPs to be predicted as rare SNPs are more difficult to tag using the common SNP panel and even then large numbers of samples are required in the reference panel to ensure adequate representation of the minor allele. Equally, imputation accuracy is reduced if there are differences in diversity or allele frequency between the reference panel and the study samples population (Marchini and Howie 2010). The completion of the 1000 genomes project will facilitate the imputation of additional common variants from GWA genotyping data, as it produces a finer scale map of genetic variants down to 1% allele frequency and will provide a much larger reference panel which will improve imputation quality at less common SNPs compared to HapMap.

Imputation has now been utilised in a number of studies and was successfully used recently to refine the association signal in a GWA study on smoking quantity (Liu *et al.* 2010). Analysis of the imputed data often identifies more strongly associated SNPs than those that were genotyped in the study.

1.10 The aims of this thesis

The main aim of this thesis is to uncover susceptibility alleles that influence the risk of colorectal cancer. The wealth of data produced from a GWA study allows for an exploration of the statistical genetic methods available for analysis of genetic data, in addition to the basic association analysis. In the search for susceptibility alleles for CRC, this thesis covers the application of such methods to six case/control datasets from our colorectal cancer GWA study and includes:

- GWA association analysis and subsequent meta-analysis for the detection of common low penetrance alleles that influence colorectal cancer risk in the population
- The fine-mapping of association signals through imputation, additional genotyping and meta-analysis of SNPs not genotyped as part of the GWA study and through gene screening to identify causal variants.
- Association analysis of the X chromosome to uncover additional variants
- An investigation into the effect of runs of homozygosity on CRC risk

- The search for moderate penetrance susceptibility alleles in single families through Linkage analysis and the study of somatic alterations, by loss of heterozygosity (LOH) analysis, in tumours from affected individuals, which may identify possible tumour suppressor genes.

Chapter 2. Materials and Methods

2.1 The GWA study

The CRC GWA study was a multistage study, consisting of a discovery stage (phase 1), a validation stage (phase 2) and a replication phase. This study was undertaken in parallel to a study of the same design, which was led by Professor Malcolm Dunlop in Edinburgh. Although our initial analyses were performed using just the London samples, we later included the Scotland GWA study in a large meta-analysis.

In phase 1, EngP1 samples were genotyped for 555,352 SNPs on the Illumina HumanHap 550k TagSNP genotyping array and ScotP1 samples were genotyped for 555,510 SNPs on the combined Illumina HumanHap300 and HumanHap240S SNP arrays. The phase 2 samples were genotyped for a subset of these SNPs that included the 14,982 SNPs most strongly associated with colorectal neoplasia in EngP1 and the 14,972 SNPs from ScotP1 (432 of these SNPs were included in both lists). Additionally 13,186 SNPs were included from a combined analysis of both EngP1 and ScotP1, and a number of candidate SNPs. In total, a panel of 42,708 SNPs were genotyped in EngP2 and ScotP2. Associated SNPs from phase 1 that were validated in phase 2 were then genotyped in additional independent cohorts in the replication phase to confirm the association. The overall design of the study is shown in figure 2.1 and the datasets are discussed in detail in the Appendix and summarised in the table 2.1.

Figure 2.1 The overall design of the GWA study

The figure below provides an outline of the case control cohorts used at different stages of the analysis in the GWA study, including the phase one and two samples to the cohorts used in the replication phase. This figure is reappears in Chapter 2 and 3 to indicate the cohorts used for the particular analysis.

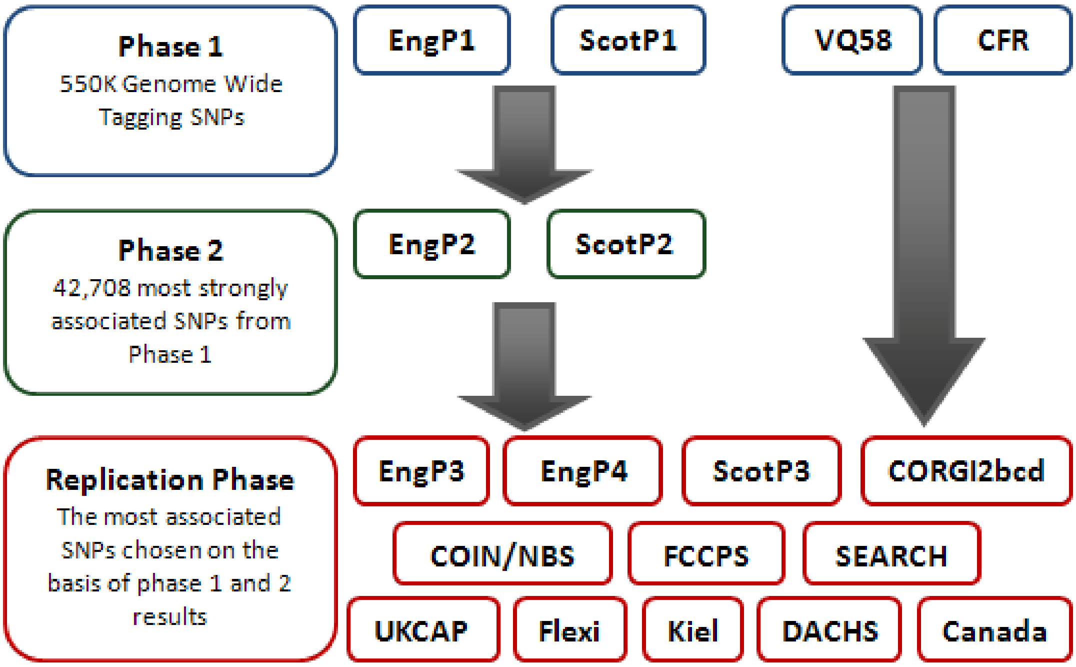

Phase 1
550K Genome Wide Tagging SNPs
EngP1
ScotP1
VQ58
CFR
Phase 2
42,708 most strongly associated SNPs from Phase 1
EngP2
ScotP2
Replication Phase
The most associated SNPs chosen on the basis of phase 1 and 2 results
EngP3
EngP4
ScotP3
CORGI2bcd
COIN/NBS
FCCPS
SEARCH
UKCAP
Flexi
Kiel
DACHS
Canada

Table 2.1 Summary of the datasets utilised in the GWA study

The numbers of male and female samples in each dataset is provided where data was available. The replication phases were only genotyped for SNPs that were taken forward for replication and so a genotyping platform is not given.

Cohort	Sample Size		Male/Female proportion		Genotyping Platform
	Cases	Controls	Cases	Controls	
Phase One Datasets					
EngP1	930	965	419/511	434/531	Illumina Hap550
ScotP1	1,012	1,012	518/494	518/494	Illumina Hap550
Phase Two Datasets					
EngP2	2,873	2,871	1,199/1,674	1,164/1,707	42,708 Illumina custom array
ScotP2	2,057	2,111	1,249/808	1,257/854	42,708 Illumina custom array
Additional GWA study datasets					
VQ58	1,432	2,697	896/536	1391/1306	Illumina Hap300
CFR	1,186	998	616/570	477/521	Illumina Hap1M

Replication datasets					
COIN/NBS	2,182	2,501	-	-	-
EngP3	3,286	3,017	2,158/1,128	1,212/1,805	-
CORGI2bcd	588	1,092	-	-	-
EngP4	1070	415	-	-	-
SEARCH	2,222	2,262	1,278/944	949/1,313	-
FCCPS	962	846	-	-	-
DACHS	1,373	1,480	790/583	719/761	-
Kiel	2,169	2,145	-	-	-
Canada	1,175	1,184	503/672	667/517	-
Replication phase used for Tomlinson et al. 2008					
POPGENSHIP	2,569	2,699	1,382/1,187	1,296/1,395	-
DFCCS	783	664	370/413	251/413	-
MCCS	515	709	270/245	352/357	-
EPICOLON	515	515	305/210	290/225	-

2.1.1 The analysis of the GWA study

Genome-wide association analysis involves comparing the allele frequency of SNPs between cases and controls to identify SNPs significantly associated with disease and determine the size and direction of the effect. If the null hypothesis of no association is true, then the allele frequencies should be approximately equal. The association analysis was performed initially in R and later using PLINK (using the –assoc and –model commands) by performing an allelic chi square test on the allele counts of each SNP in cases and controls (in a 2x2 table) to produce a P value to determine whether there was a significant difference in allele frequency. The effect size and direction is determined using the odds ratio (OR), which gives an approximation of the relative risk of developing the disease in an individual carrying the risk genotype. The 95% confidence intervals (CI) for the OR were also calculated.

For imputed data, the association analysis was performed in SNPTEST, using tests designed to take the uncertainty of the imputed genotype probabilities into account that are discussed in Chapter 4. For these analyses, beta (log OR) and standard error (SE) values were presented instead of ORs.

2.1.2 Meta-analysis of Genome-Wide Association Results

Meta analysis for genotyped SNPs was performed in R using the Mantel-Haenszel method for combining results, which provides a combined odds ratio (OR) across studies, under both a fixed and random effects model. The scripts used are provided in the appendix. The random effects model assumes that each dataset can have a different genetic effect and is better suited to deal with between study heterogeneity, although this will result in a wider confidence interval for the OR.

In this study, the analysis was performed under both fixed effects and random effects models. The P value generated under a fixed effects model was used unless there was evidence of between-study heterogeneity. The presence of heterogeneity between studies will cause a large difference between the P values obtained from each model and in these instances the random effects model P value was used.

Another, more robust, method to detect heterogeneity is to use a heterogeneity score such as I^2, which is the percentage of total variation across studies caused by heterogeneity or Cochran's Q statistic to test for between-study heterogeneity (P heterogeneity). These were both presented with the meta-analysis results in Chapter 4 and 5, where SNPs were rejected if the P value for between-study heterogeneity was below 0.05.

Meta-analysis for imputed data was performed using the program META by combining P values from the SNPTEST output (frequentist additive score test) across datasets to avoid converting imputed genotype probabilities into genotype counts, the reasons for this are discussed in Chapter 4.2.4

The R script for meta-analysis that was used to combine the association results of multiple datasets for the genotyped SNPs was written by Emily Webb, while a member of Richard Houlston's group at the ICR in Sutton. I formatted the association analysis results correctly using an awk script, which is given in the Appendix section 9.2, along with the meta-analysis R script.

2.1.3 Logistic regression to determine independent effects

Regression analysis is a method for investigating the relationship between a dependent (response) variable and independent (predictor) variables, possibly while taking account of other independent variables or covariates, by attempting to fit the data to a model. In logistic regression, the response variable is a binary variable. A logistic curve is fit to the independent variables count values to determine the goodness of fit and produce a regression P value.

Logistic regression analysis was performed in PLINKv1.07 using the '--logistic' and '--condition' commands (http://pngu.mgh.harvard.edu/~purcell/plink/index.shtml).

2.2 Prediction of untyped genotypes using IMPUTE

2.2.1 The imputation basic work flow

After quality control of the genotype data, the genotypes were converted from PLINK format into the IMPUTE format using the program GTOOL (which is available at the following website: http://www.well.ox.ac.uk/~cfreeman/software/gwas/gtool.html).

Two different versions of IMPUTE were used in this study, version 1 (for imputation of HapMap SNPs and the X chromosome SNPs) and version 2 (for imputation using more than one reference panel), from https://mathgen.stats.ox.ac.uk/impute/impute.html. Only SNPs present in the reference files were included in the IMPUTE format genotype files. As the positions of the SNPs in EngP1, ScotP1 and VQ58 were calculated using the NCBI35 genome build and the HapMap phase two reference panel SNPs were NCBI36, I converted the positions to match the reference panel and re-sorted the SNPs by position, using a small awk script (change_locs.awk), given in the Appendix section 9.3, prior to analysis.

I then separated each chromosome into 7Mb segments, created a command file for each segment and produced a script to run IMPUTE across all segments for each chromosome in a loop. The shell scripts are given in the Appendix section 9.3.2. The size of each chromosome was retrieved from the UCSC website using the NCBI36 genome build.

The command line for each segment was thus:

```
chr1_1.com

cd /farm/home/spain01/projects/GWA/impute ; ./impute
-h b36_files/hapmap_r24_b36_fwd.consensus.qc.poly.chr1_ceu.phased
-l b36_files/chr1.ceu.r24.legend -m b36_files/genetic_map_chr1_CEU_b36.txt
-s hapmap_impute1/chr1_VQv2_hap.strand -g
hapmap_impute1/VQv2c_chr1.hap36.gen
-Ne 11418 -o hapmap_impute1/chr1/VQv2c_chr1_1.imputed
-i hapmap_impute1/chr1/VQv2c_chr1_1.info -r
hapmap_impute1/chr1/chr1_1.2summary
-int 0 7000000
```

2.2.1.1 The X chromosome

The imputation of SNPs on the X chromosome was performed using IMPUTEv1 with the following commands. For the X chromosome a sample file must be included at this point to determine the gender and process the male and female samples appropriately.

```
VQ_X_1.com
cd /farm/home/spain01/projects/GWA/impute/; ./impute -chrX -h
chrX_files/genotypes_chrX_CEU_r21_nr_fwd_non-par_phased_by_snp_no_mono
-l chrX_files/genotypes_chrX_CEU_r21_nr_fwd_non-par_legend.txt
-m chrX_files/genetic_map_chrX_non-par.txt -fix_strand
-g hapmap_impute1/VQv2clean_chrX.hap35.gen
-sample hapmap_impute1/VQv2clean_chrX.sample_2 -Ne 11400
-o hapmap_impute1/chrX/VQv2clean.chrX_1_aff.imputed
-i hapmap_impute1/chrX/VQv2clean.chrX_1_aff.info
–r hapmap_impute1/chrX/chrX_1.VQv2c.summary -int 0 7000000
```

2.2.1.2 Imputation with multiple reference panels

For imputation performed using IMPUTEv2, there were multiple reference panels and the commands were slightly different. The command file below is for the imputation of VQ58 using the EngP1 and HapMap2 reference panels:

```
chr1_1.com
cd
/farm/home/spain01/projects/GWA/impute/impute2/impute_v2.1.0_x86_64_static;
./impute2 -h ../b36_files/hapmap_r24_b36_fwd.consensus.qc.poly.chr1_ceu.phased
-l ../b36_files/chr1.ceu.r24.legend -m ../b36_files/genetic_map_chr1_CEU_b36.txt -
g_ref ../engp1_data/v5/b36_genfiles/p1_chr1.ctrl.hap36.gensort -fix_strand_g_ref  -
fix_strand_g
-g ../hapmap_impute1/VQv2c_chr1.hap36.gen -Ne 11418  -k 80 -iter 30 -burnin 10 -o
../engp1hap2_impute/chr1/VQv2c_chr1_1.imputed
-i ../engp1hap2_impute/chr1/VQv2c_chr1_1.info
-r ../engp1hap2_impute/chr1/chr1_1.summary -int 0 7000000
```

Instead of creating a strand file, I utilised the flip-strand function in IMPUTE to ensure that the strand matched the reference panel. The genotyped SNPs are distinguished from the imputed SNPs by the SNP type. In IMPUTEv1 imputed SNPs are labelled as '---', while imputed SNPs have a chromosome ID. IMPUTEv2 labels SNPs type 0 if the SNPs are only in the phased reference panel, type 1 if they are only in an unphased reference panel, type 2 if they are genotyped in the study (or inference) panel and also in one of the reference panels.

2.2.2 SNPTEST

A SNPTEST command file was produced for each imputed segment and the scripts used to produce and run these commands are given in the Appendix section 9.3.3.

The results of this analysis were then pruned to filter out SNPs that did not pass the criteria based on the information score greater than 0.5, number of samples with a maximum genotype probability less than 0.9 and the minor allele frequency.

The meta-analysis of imputed data was performed using the program META (http://www.stats.ox.ac.uk/~marchini/software/gwas/gwas.html), which accepts the output of SNPTEST, this allowed me to use the P values with effect sizes and directions as calculated using the 'proper' score test method, rather than the best guess genotype counts that were used with the R method given above.

2.3 Extraction of DNA from paraffin embedded tumour samples

The following method uses the QIAamp DNeasy kit (Qiagen) and is optimised for the extraction of DNA from paraffin embedded tumour samples that is of a quality for use with the Illumina GoldenGate SNP genotyping arrays, which requires 1µg of DNA in TE solution at 50ng/µl. This experiment was performed on adenomas, which yield less DNA than cancers, but also have less heterogeneity and are more easily macro-dissected from the surrounding normal tissue. Depending on the size of the tumour, between five and ten 10 micron thick slices of tissue were cut from each paraffin block to make slides, which were stained with tol blue to better distinguish the tumour from the normal tissue.

Extraction of DNA

Briefly, for each tumour, the required tissue was scraped, using a fine needle or scalpel, from each slide into a 1.5ml eppendorf tube containing 180µl of Buffer ATL™. Then 20µl of 20mg/ml Proteinase K was added and the tube vortexed to thoroughly mix the contents. The samples were then incubated at 55°C overnight and another 20µl of Proteinase K was added in the morning and the tube vortexed again. At the end of the day, a final 20µl of Proteinase K was added and after vortexing the samples were incubated at 55°C overnight. Throughout this process the samples were mixed at regular intervals. At this stage, if the samples appeared to be fully digested, 200µl of Buffer AL™ was added to the tubes, which were then vortexed and incubated in a 70°C heat block for 10 minutes. The tubes were then briefly centrifuged, to remove drops from the lid, and 200µl of 95% Ethanol added before vortexing to mix.

Purification of DNA

The DNA in the sample is purified by passing the contents of the tube through a fast spin column, which is placed within a 2ml collection tube. The QIAamp silica-gel membrane in the spin column specifically binds DNA, while any contaminating material in the mixture is eluted into the collection tube. Multiple washes are performed to remove proteins and other substances to yield pure DNA. After adding the sample to the top of the spin column, without wetting the rim, the lid is closed and the tube centrifuged at 8000rpm for 1 minute. The collection tube is discarded and the spin column placed in a new collection tube.

500µl of Buffer AW1™ is then added to the spin column, which is centrifuged at 8000rpm for 1 minute and the filtrate discarded. 500µl of Buffer AW2™ is then added to the spin column, which is centrifuged for 1 minute at 13000rpm and the filtrate discarded. To remove any residual buffer the column was inserted into a new collection tube and centrifuged at 13000rpm again for 1 minute.

Elution of DNA

The DNA bound to the spin column membrane is finally eluted by the addition of 100µl of Buffer AE™ to the column, which has been inserted into a 1.5ml Eppendorf tube and is then centrifuged at 13000rpm for 1 minute. The volume of the Buffer AE™ added was reduced to 50µl for very small adenomas. To increase the final yield this step was repeated by applying the filtrate back into the spin column and centrifuging again. The concentration of the DNA sample, in the Eppendorf tube, was then measured.

2.4 The quantification of DNA

The concentration of extracted DNA was determined by spectrometry using 1µl of DNA solution in a SPECTRAmax PLUS ('Nanodrop') spectrophotometer to measure optical density (OD) at 260nm and 280nm. The concentration of the DNA in µg/ml is then calculated by 'dilution factor x 50 x OD_{260}', based on the knowledge that an OD_{260} of 1 is equal to 50µg/ml of double stranded DNA. The quality of the DNA can be assessed using the ratio of OD_{260}/OD_{280}, where a value of 1.8 should be achieved for a pure DNA sample.

2.5 Standard PCR protocol

The polymerase chain reaction (PCR) consists of many cycles of three steps: denaturation of the DNA by heating to 95°C to separate the strands, annealing of the primers to the separated DNA strands at a temperature specific to the primers (50-70°C) and then amplification of the product by DNA synthesis at about 70°C. The DNA synthesis occurs through the addition of a DNA polymerase and deoxyribonucleoside triphosphates (dNTPs), one for each base, to the reaction mix. A standard 25µl reaction consisted of 1µl DNA (20ng/µl), 2.5µl PCR buffer (Promega), 1.5µl $MgCl_2$ (Promega), 2µl dNTPs (Amersham), 0.25µl Taq DNA polymerase, 0.25µl of both forward and reverse oligonucleotide primers (Sigma) and 17.25µl ddH_20.

Cycling conditions for the AR55 program were 5 minutes at 94°C, 35 cycles of denaturing at 95°C, annealing at 55°C and amplification at 72°C with one minute at each temperature, followed by a final 10 minutes at 72°C. I then used gel electrophoresis, with a 2% agarose gel, to ensure that the PCR product was the expected size and to check the success of the reaction.

All primers used for this thesis are given in the Appendix with the associated PCR conditions, such as $MgCl_2$ concentration and annealing temperature. The primers for gene screening, by sequencing or the LightScanner, were designed using a combination of ExonPrimer, which is accessed through a link on the UCSC genome browser (http://genome.ucsc.edu/) and the primer3 program (available at http://frodo.wi.mit.edu/primer3/). Primers for SNP genotyping were designed using the KBioscience ARMS primer design tool, “primer picker”, although this has since been removed from the website and replaced with a design service (http://www.kbioscience.co.uk).

2.6 Agarose Gel Electrophoresis

Gel electrophoresis is the process where molecules are separated owing to their size by moving through a porous gel matrix charged with an electric current. As DNA molecules are negatively charged, the molecules move away from the negative electrode in the direction of the positive electrode. The speed that the molecule travels through the gel is size dependent and smaller DNA molecules will travel faster than large ones.

A 2% agarose gel is made by dissolving 20g of agarose into 1L of 1xTBE solution in a 2 litre beaker. The solution is heated in a microwave to fully dissolve the agarose until the mixture became clear. When the solution has cooled slightly, Ethidium Bromide (at a final concentration of 0.25µg/µl) is added to the agarose gel to allow visualisation of the DNA bands due to fluorescence under ultra violet light. The gels are poured into tray moulds to set and then placed in the electrophoresis tank, covered with 1xTBE buffer. 5µl of DNA was added to 3µl of loading dye (Orange G, Trevigen) add then loaded to the wells. An additional sample of 7µl of 1Kb ladder was also loaded to the gel to provide a size reference (Gibco BRL). Generally, PCR products were run on a 2% agarose gel at 120V for 20mins connected to a DC power pack (Bio-rad PowerPac 300).

2.7 Fluorescent DNA sequencing protocol

After amplifying the desired DNA region by PCR (using primer concentrations of 20µM), and running a sample of the product on an electrophoresis gel to determine that the PCR product is of the expected size and to gain an indication of the amount of DNA present, preparation for Sanger sequencing can begin. The PCR products were first purified, to remove excess dNTPs and primers from the PCR reaction, by adding 5µl of each sample to a new 96 well plate and then adding 2µl of ExoSAP-IT® (USB). The plate is then placed on a Tetrad thermal cycler (MJ Research) for 15 minutes at 37°C and then 15 minutes at 80°C to inactivate the ExoSAP-IT enzymes (Exonuclease 1, to remove primers, and Shrimp alkaline phosphatase, to remove excess dNTPs).

Depending on the strength of the DNA band on the electrophoresis gel, the PCR products were then diluted by 20µl if the band was very strong and 10µl for a weak band. The diluted PCR products were then used in a sequencing reaction, where 4µl of diluted PCR product was added to 8µl of BigDye terminator ready reaction mix (BDT, PE Applied Biosysyems), 1µl of forward primer (at a concentration of 2µM) and 7µl of ddH_20. This reaction is also performed with the reverse primer to obtain the sequence for the reverse strand. The sample is then placed on a thermal cycler with the following conditions: 4 minutes of initial denaturation, then 25 cycles of 30 seconds at 94°C, 10 seconds at 50°C and 4 minutes at 60°C, followed by a final 7 minutes at 60°C.

The DNA samples were then purified using the DyeEx 2.0 spin kit (Qiagen) following the manufacturer's instructions, to remove any excess dye terminator BDT. This essentially involves eluting the 20µl sample from the sequencing reaction through a spin column (although 96 well plate versions were also used) into a 1.5ml Eppendorf tube by centrifuging at 3000rpm for 3 minutes. The filtrate can then be used for sequencing.

The DNA samples were denatured by incubating at 94°C for 4 minutes and then sequenced by running on a 5% polyacrylamide gel using the semi-automated ABI Prism 377 XL sequencer (PE- Applied Biosystems). Sequence analysis was performed using the Base Calling option of the Sequencing Analysis Programme (Version 2.1, PE-Applied Biosystems). Sequences were read using FinchTV

(http://www.geospiza.com/Products/finchtv.shtml) and compared using the Seqman program, which is part of the DNASTAR lasergene 8 package.

2.8 SNP genotyping using KBioscience allele specific PCR (KASPar)

This method uses 3 primers, one forward primer specific to the common allele of the SNP and one primer specific to the alternative allele and a common reverse primer (all at a concentration of 200µM). The KASP reaction mix contains a passive reference dye, ROX. The allele specific primers are labelled with different fluorescent dyes, either FAM or VIC.

The basic protocol, performed to the manufacturer's instructions, involved producing a 500µl Assay mix specific to each SNP, which was made up of 30µl of the common allele primer, 30µl of the alternative allele primer, 75µl of common primer and 365µl of

ddH_20. For each sample, 1µl of DNA was added to genotyping mix, consisting of 0.165µl of Assay mix, 3µl of 4xKASP reaction mix, 0.039µl of Taq DNA polymerase enzyme (Promega), 0.096µl of 50mM $MgCl_2$ and 7.7µl of ddH_20, to produce a 12µl reaction. A mastermix was created of the genotyping mix for at least 210 samples, to avoid pipetting small volumes. The mixture was kept on ice and 11µl of the mastermix was added to each well of a PCR 96 well microtitre plate and then 1µl of sample DNA added.

The plate was then sealed with an optically clear adhesive lid and placed on a thermocyler for the following conditions: 94°C for 15 minutes (for hot start enzyme activation), followed by 20 cycles of 94°C for 10 seconds, 57°C for 5 seconds and 72°C for 10 seconds, followed by 18 cycles of 94°C for 10 seconds, 57°C for 20 seconds and 72°C for 40 seconds. The plate was then briefly centrifuged and read using a fluorescent plate reader and the data plotted with FAM against VIC to determine genotypes.

2.9 Mutation screening and SNP genotyping using the LightScanner

The LightScanner (Idaho Technology Inc.) is a mutation scanning and genotyping method that uses High Resolution Melting analysis of PCR products. This technique allows mutation detection through the comparison of DNA melting curves between the DNA sample and a control reference. The process relies on a dye called LCGreen Plus (Idaho Technology Inc), which is a fluorescent dye that binds to double-stranded DNA, and is incorporated into the DNA by PCR at the initial amplification stage. LCGreen is also able distinguish between heterozygotes and homozygotes in a sample and was thus used to scan for mutations in stretches of DNA and for SNP genotyping (with a reference sample added for each possible genotype).

The master mix for the LightScanner PCR, for multiple samples, was similar to a standard PCR reaction and, for each sample, consisted of 1.25µl PCR buffer (Promega), 0.75µl $MgCl_2$ (Promega), 1µl dNTPs (Amersham), 0.125µl Taq DNA polymerase, 0.125µl of each 20µM oligonucleotide primer (Sigma), 2.5µl Q-Solution (Qiagen), 4.375µl of ddH_20, and 1.25µl of LCGreen Plus dye. Q-Solution changes the melting properties of DNA and can improve the quality of PCR products. An 11.5µl aliquot of the master mix was pipetted into an opaque white 96 deep well plate and covered with 20ml of mineral oil to prevent any evaporation from the well. After the addition of 1ml of DNA to the bottom of each well, the plate was covered with a self adhesive lid.

The amplification of DNA was performed as per PCR, except that the final 94°C denaturation step lasts for 30 seconds and is followed by a 20°C holding temperature.

Before reading the plate in the LightScanner, the contents were briefly spun in a centrifuge to remove any droplets from the lid and avoid contamination when the lid is removed prior to inserting into the LightScanner for scanning and melting analysis. After reading the results in the LightScanner, the amplification products were purified to remove primers and surplus dNTPs and the DNA used in a subsequent sequencing reaction to confirm any variants that were discovered. The physical positions, based on the NCBI genome build 36, of any identified variants were determined by applying the BLAT function to each exon sequence, available on the UCSC website (http://genome.ucsc.edu/).

2.10 Solutions

PBS (Phosphate Buffered Solution): 8g NaCl (BDH), 0.2g KCl, 1.44g Na_2HPO_4, 0.24g KH_2PO_4, 800ml dH_2O. Adjusted to pH 7.2 using HCl and made up to 1L with dH_20.

TE (Tris EDTA) Buffer: 10mM Tris, 1mM EDTA, pH 7.5

Proteinase K (stored at -20°C): 2mg proteinase K (BDH) and 100ml acetic acid (BDH)

TBE (Tris Borate EDTA) Buffer: 1L of 5x solution use 53g Tris base, 27.5g Boric Acid and 20ml of 0.5M EDTA (pH 8)

2.11 Homozygosity mapping using PLINK

Here follows the basic workflow for the analysis of homozygosity mapping in cases and controls from the EngP1 cohorts, which is described in Chapter 6. Statistical analysis for this study was performed using R (version 2.7.0).

2.11.1 Analysis of homozygosity by SNP

The counts for each sample were generated using a personal script, chisq_multi.R given in the Appendix in section 9.4.1, which performs the analysis using a standard 2x2 χ^2 test and calculates ORs in R. I compared the counts of individuals homozygous for either allele against the counts for heterozygotes, for each SNP, in cases and controls. The input files, corgi.aff.counts and corgi.ctrl.counts, were generated from the output files of the PLINK 'model' association analysis.

2.11.2 Meta-analysis of homozygosity association results

Meta-analysis was performed using the meta package in R using the Mantel-Haenszel method for combining results under both fixed and random effects models. The script

‘2grpmeta.R’ is given in the Appendix section 9.4.2, and was modified from an original script provided by a colleague, Emily Webb.

2.11.3 Analysis of recurrent ROH regions and comparison with detected CNVs

By applying the option ‘homozyg-group’ to the command line when running the ROH tool, an additional output file, plink.hom.overlap, is produced that contains lists of the ROHs that overlap with those detected in other individuals separated into pools. Each pool is separated by a row labelled CON, which contains the consensus region covered by the ROHs in the pool, the number of cases and controls carrying an overlapping ROH, start and end positions and the size of the region in SNPs and kilobases. An ROH was considered to be recurrent if it was present in more than five individuals. This data was used to perform an analysis comparing the number of cases and controls with an ROH overlapping the consensus region. The rows containing the consensus region details were placed in a separate file and the P values calculated using the R script, *ROH_Pvalue.R* given in the Appendix section 9.4.3, if the cell count was less than 5, Fishers exact test was used instead of the Chi Square test.

2.12 Linkage Analysis

The single family linkage analysis was performed using Allegro (Gudbjartsson *et al.* 2005) with the same parameters as previously published (Kemp *et al.* 2006). Allegro is a fast multipoint linkage analysis program, which is efficient over large numbers of SNPs.

The individuals were assigned one of four liability classes depending on the age of diagnosis, which was based on a segregation analysis on CRC families (Aaltonen *et al.* 2007). The penetrances used in this analysis are given in Table 2.3. Individuals that met the CORGI criteria with adenomas, instead of cancer, were considered to be equivalent to CRC 15 years later and so the age of diagnosis was adjusted to reflect this. This correction was based on published data on the estimation of malignant transformation from adenoma to cancer (Chen *et al.* 2003).

Table 2.3 The parametric analysis liability classes with penetrances and phenocopy rates for the dominant and recessive model analyses

Class	Age	Dominant Model			Recessive Model		
		Zero alleles (phenocopy)	One allele	Two alleles	Zero alleles (phenocopy)	One allele (phenocopy)	Two alleles
1	<50	0.0004	0.044	0.044	0.00004	0.00004	0.054
2	50-59	0.002	0.105	0.105	0.0003	0.0003	0.146
3	60-69	0.007	0.213	0.213	0.0026	0.0026	0.331
4	>70	0.07	0.42	0.42	0.06	0.06	0.638

The SNPs used in this analysis were from the Genechip Mapping 10K Xba 142 SNP array (Affymetrix Inc. Santa Clara, CA) consisting of approximately 10,000 SNPs. In order to prevent any confounding of the results caused by LD relationships between SNPs, the dataset was pruned to remove SNPs in pairwise LD (in this case defined as r^2 greater than 0.16). SNPs were also removed for Mendelian errors based on impossible genotypes. Overall, this resulted in a total of 7,228 SNPs that were included in the analysis.

The allele frequencies used for these analyses were estimated from all individuals that were genotyped as part of the study, not just those that were included in the single family analysis. This was done to try to ensure that the allele frequencies used closely represented the sample population, which may not be the case if using the allele frequencies in HapMap.

The analyses in Allegro version 2.0f (Gudbjartsson *et al.* 2005) were performed using the following models with the founder couples option activated:

Parametric analysis: MODEL mpt par het

Non-parametric analysis: MODEL mpt exp pairs equal, MODEL mpt lin pairs equal, MODEL mpt exp all equal and MODEL mpt lin all equal.

I performed a multipoint analysis where all markers were analysed simultaneously and the LOD score of one marker also takes into account surrounding markers. The parametric models were defined using the parameters detailed above for the dominant and recessive models (see Table 2.3). For the non-parametric analysis, an allele sharing model (linear and exponential) was used. The 'pairs' and 'all' options call the S_{pairs} and S_{all} scoring functions, respectively, for determining the level of shared alleles IBD. S_{pairs} is calculated using the number of alleles that are shared IBD between pairs of affected relatives and S_{all} used the number of alleles shared IBD between all affected members of the family (McPeek 1999).

To generate estimated maximum LOD scores attainable for each family, simulations were performed in Allegro version 1.2c assuming a single gene parametric disease model and using the pre and dat files created for the dominant or recessive models with the following parameters: SIMULATE dloc:50.53 npre:1 rep:1000 err:0 yield:1 het:0. The values for error rate (err), yield and heterogeneity (het) are the default values. This produces 1000 sets of simulated genotypes for each of the genotyped members of each family based on the allele frequency, pedigree structure, disease model and individual phenotype. These genotypes were then used in a recessive and dominant linkage analysis using the parameters described above.

2.13 The detection and analysis of segments shared IBD

PLINK contains a set of functions to detect segmental sharing between samples (http://pngu.mgh.harvard.edu/~purcell/plink/). The first step uses the '--genome' function to determine the relatedness between individuals using IBS, which is used subsequently in the analysis. Shared segments are then detected with the '--segment' option using the previously created genome file and the individual genotypes as input files. The output provides a 'segment' file, which lists all shared regions between individuals with sample ID, physical position and size of region, an '.indiv' file providing the total size of shared regions between any two individuals and an 'overlap' file, which groups each shared region that is found in more than one sample into pools and lists the number of cases and controls included in each pool. The final step is to perform an analysis of the shared segments IBD to identify segments that are shared more often between cases than controls. This is performed in PLINK by using the '--mperm' option with the 'segment' file (containing the pairs of shared segments) as the input. The actual command line instructions used to run this analysis was as follows:

Step 1: Determine segment sharing by IBS

Calculate segment sharing by IBS using a SNP panel pruned for pair-wise LD (r^2>0.2), which produces the file corgiPCA_50502_g.genome (The output was then filtered to only include pairs of individuals with a PI-HAT score between 0 and 1):

```
./plink --noweb --bfile corgiPCA_50502hapb36 --genome --out corgiPCA_50502_g
```

Step 2: Infer shared segments IBD

I used the segment function, which produces the file corgiPCA_50502_g.seg.segment:

```
./plink --noweb --bfile corgiPCA_50502hapb36 --read-genome
corgiPCA_50502_g.genome
--segment --cm --segment-group --out corgiPCA_50502_g.seg
```

The default settings were used to select the segments (minimum segment length of 1000kb and 100 SNPs). This restricts the analysis to those larger segments that more likely to be shared IBD.

Step 3: Statistical analysis of shared segments

Perform the analysis to determine whether there is a statistically significant higher rate of sharing in case/case sample pairs than non-case (or discordant) sample pairs:

```
./plink --noweb --bfile corgiPCA_50502hapb36 --read-segment
corgiPCA_50502_g.seg.segment --cm --mperm 10000 --out corgiPCA_50502_g.assoc
```

Chapter 3. Detection of susceptibility alleles for CRC by genome-wide association analysis

3.1 Introduction

Until a few years ago, the known genetic and hereditary components for CRC were limited to rare, highly penetrant dominant cancer conditions, such as FAP and HNPCC, which explained just 5% of the variation in risk. However, it was expected that 35% of the variation in risk could be explained by a heritable factor (Lichtenstein *et al.* 2000). The advances in the knowledge of SNPs, LD patterns and the technology to genotype thousands of samples for hundreds of thousands of SNPs, gave us the ability to study the frequency of common variants in cases and controls and determine whether these can confer disease susceptibility. This was the basis of the 'common disease common variant' hypothesis (Lander 1996). The idea suggested that there existed a number of variants at various loci that are common in the population and that individually have a small effect, but in combination can confer a greatly increased susceptibility to common diseases (Hardy and Singleton 2009).

The detection of these susceptibility variants indirectly by GWA studies relies on the assumption that the disease loci are not heterogeneous. If there were multiple low frequency susceptibility variants within a gene, or locus, then the power to detect them by association mapping would diminish greatly (Slager *et al.* 2000).

This chapter covers aspects of the GWA study that were conducted as a group that was part of a large collaboration and provides a summary of the variants discovered. I describe the design and progression of our GWA study for the detection of common susceptibility variants for CRC and my personal input in the aspects of data analysis, quality control, replication genotyping and meta-analysis with additional datasets. This work is expanded in the following chapter, where I discuss candidate gene screening and explore the use of imputation to predict genotypes of untyped SNPs for the fine-mapping of associated hits and better overlap of SNPs between datasets thus facilitating meta-analyses to identify new associated variants.

3.2 GWA Study Design

The aim of this study was to identify common (minor allele frequency (MAF) greater than 1%) low penetrance variants that influence CRC risk by comparing allele frequency between cases and controls in a large GWA study. The sample datasets used in this study are described in detail in the Materials and Methods.

In order to remove the need to genotype all samples for the maximum number of SNPs, but maintain the power to detect an association, we conducted a multi-stage analysis, as described in Chapter 1 and 2. The cases from EngP1 all had at least one first degree relative with CRC to maximise the power of the initial discovery phase by increasing the chance that the affected individual carries the susceptibility allele. The controls were also free of personal or familial history of CRC. Compared to GWA studies of cases with no family history, this reduces the number of samples required to detect a risk allele and allowed us to genotype a smaller enriched dataset in the first phase for the maximum number of SNPs (illustrated in Figure 1.4 in Section 1.8.4). The validation phase of the study involved genotyping a second much larger dataset

(EngP2) for the most associated SNPs that were identified in the first phase and then combining the results in a meta-analysis to give a single test statistic. This approach is the most efficient and cost effective (Skol *et al.* 2006).

Our GWA study was performed in parallel with another study, of the same multi-stage design and genotyped for the same SNPs, led by Professor Malcolm Dunlop in Edinburgh (Tenesa *et al.* 2008). The phase one and phase two datasets from this study are referred to as ScotP1 and ScotP2, respectively. The EngP2 and ScotP2 datasets were genotyped for the same SNPs, facilitating eventual meta-analysis (see Section 2.1). The most strongly associated SNPs selected from the joint analysis of the EngP1 and EngP2 datasets, were genotyped in the replication phase consisting of additional datasets ascertained by ourselves, our main collaborator Professor Richard Houlston of the Institute of Cancer Research and additional groups that together form the Colorectal Cancer Genetics consortium (COGENT)(Tomlinson *et al.* 2010).

The phase one datasets were genotyped on the HumanHap550 SNP array. The tagging SNPs on this array efficiently tag 80% of common variants in HapMap that have a minor allele frequency (MAF) of more than 20%. However, the array has low power to detect variants with MAF less than 10%.

3.2.1 Meta analysis

As the datasets used in this study were all recruited separately from different centres, the samples were not simply merged and analysed together. To overcome any heterogeneity between datasets, the evidence obtained in each dataset was jointly analysed in a meta-analysis. This can be done either by combining P values or by combining the effect sizes. Meta analysis was performed using the Mantel-Haenszel method for combining results, which provides a combined odds ratio (OR) across studies, under both a fixed and random effects models.

3.3 Quality control of GWA study datasets

Quality control of the data included basic tests of genotyping quality, where samples were excluded if the overall call rate was less than 95% and SNPs were rejected if less than 95% of samples were successfully genotyped or if the controls failed Hardy-Weinberg Equilibrium (HWE), as determined by Chi square test. Tests were also performed to detect population stratification, discussed below, and we calculated the genomic control inflation factor (λ) to test the level of inflation of the association test statistics in each dataset. None of the datasets showed any significant inflation of the association test statistics. In EngP1, 3 cases and 25 controls were removed, as a result of the inclusion of related samples, that were detected through an analysis of alleles identical by state (IBS), and 63 owing to changes to phenotype status that were not known at the time of genotyping. As a result of these analyses, a number of SNPs and samples were removed from the analysis (see Table 3.1). Samples and SNPs were also removed from the 1958 birth cohort based on the suggested exclusion lists provided with the data based on various quality controls including plate effects.

Table 3.1 Summary of rejected samples and SNPs per dataset

For VQ58 the number of SNPs removed includes those that were rejected, for reasons not restricted to genotyping failure, from the 1958BC which totalled 215,732 SNPs, and included on the Hap300 chip. The final SNP numbers were determined after the removal of SNPs failing HWE at $P<1x10^{-5}$. 238 of the samples rejected from the VQ58 dataset were controls.

Dataset	λ	Genotyping Failures / rejects		SNPs failing HWE	PCA rejected samples	Total remaining		
		Samples	SNPs			Cases	Controls	SNPs
EngP1	1.03	116	34	650	62	886	902	549,140
EngP2	1.04	73 (5/68)	202	13	-	2,852	2,818	51,842
ScotP1	1.02	5	240	679	29	965	984	550,639
ScotP2	1.056	44	209	8	-	2,006	2,057	51,892
VQ58	1.02	243	18,284	263	25	1,425	2,690	292,543
CFR	1.098	0	1,938	1,840	6	1,186	998	1,007,231

3.3.1 Principal components analysis

Population stratification is caused by allele frequency differences between cases and controls that can lead to false positive associations. The most likely cause for this is the inclusion of samples from different ethnic groups, which have varying allele frequencies across SNPs. Before principal components analysis (PCA), the methods available for the detection of stratification were STRUCTURE (Pritchard *et al.* 2000) and genomic control, where test statistics are corrected on the basis of the inflation factor (λ) (Devlin and Roeder 1999). STRUCTURE clusters genotype data across multiple markers to separate individuals into different populations based on population structure. The accuracy of the assignments into clusters is dependent on the number of samples, markers and degree of differentiation. This was the technique used for the initial quality control checks to determine population stratification for this GWA study and were performed by others. The results indicated that, for each individual dataset, there was no discernable structure in the population and no detectable stratification between cases and controls that might lead to erroneous positive association results.

As the number of datasets included in the study grew, I explored the use of PCA using the program smartpca in Eigenstrat to identify structure within populations. This technique has advantages over other available methods, such as STUCTURE, as it can rapidly handle large datasets that have been genotyped for hundreds of thousands of SNPs. PCA does not separate individuals into clusters, but converts genotype data into continuous axes of variation with a small number of dimensions that can be used to describe the variability within a dataset (Price *et al.* 2006). Each sample is assigned an eigenvector for each axis with the first axis describing the most statistically significant variation. The eigenvector values can be plotted against one another to identify clusters and outliers within the dataset. Smartpca also performs formal significance tests on the differentiation within the population using the eigenvalues for each component.

It has been observed that regions of long range LD exist between markers that are some distance apart. Price and colleagues identified several such regions that can confound PCA analysis (Price *et al.* 2008) and markers within these regions were also removed. I performed a PCA analysis on each dataset individually and then combined multiple datasets together to ensure compatibility for combined analysis. The SNP panel used for each PCA analysis only contained SNPs that were present in all datasets, to ensure there was no bias caused by samples that were genotyped for different SNPs, and so the number of SNPs varied between each single dataset and the grouped analyses. Also, the analysis was performed using only SNPs that are in approximate linkage equilibrium and so the dataset was pruned using the SNP pruning function in PLINK at a threshold for r^2 of 0.2 (process described in the materials and methods).

3.3.1.1 VQ58

VQ58 was the only included dataset where the cases were genotyped in house and the controls were publically available population controls genotyped on a slightly different SNP array and I performed the quality control measures in this dataset. The association results from the analysis of VQ58 showed little evidence of inflation of the test statistics (see Figure 3.1). I performed a PCA to ensure that the cases from the VQ58 dataset (which is made up of samples collected as part of two studies: VICTOR and QUASAR2) were sufficiently matched, without signs of structure, to the WTCCC2 1958 birth cohort control samples for GWA analysis. This analysis resulted in the removal of 24 cases and 1 control from the study, this consisted of outliers with eigenvector 1 greater than -0.05 and the two outliers at eigenvector 2 of 0.1 and less than -0.1 (see Figure 3.2)

Figure 3.1 The QQ plot for VQ58

This plot includes only SNPs that passed quality control measures and does not include SNPs that were out of HWE. The results showed little evidence of over inflation of the test statistics with λ=1.018.

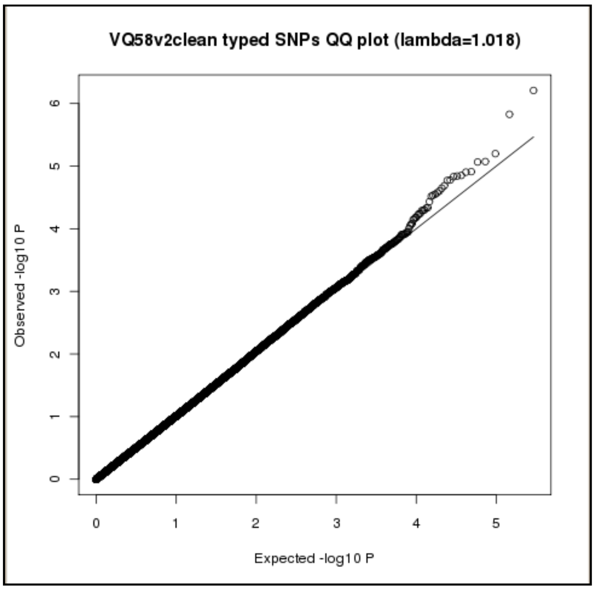
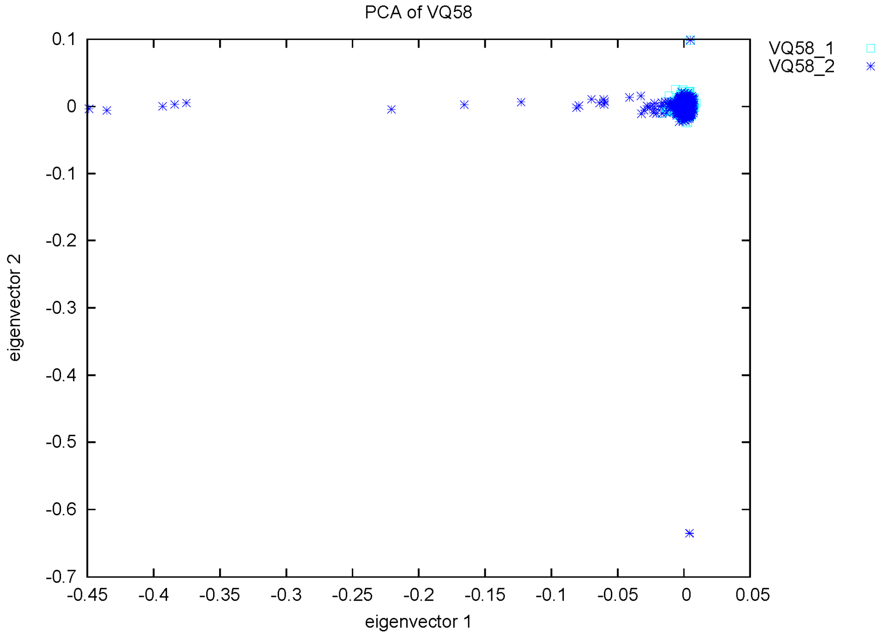


Figure 3.2 PCA results for the VQ58 dataset

The analysis was performed on 80,915 SNPs (r^2<0.2). The cases are denoted by VQ58_2 and the controls are VQ58_1. These result led to the removal of 24 cases and 1 control from the study, this consisted of outliers with eigenvector 1 greater than -0.05 and the two outliers at eigenvector 2 of 0.1 and less than -0.1.

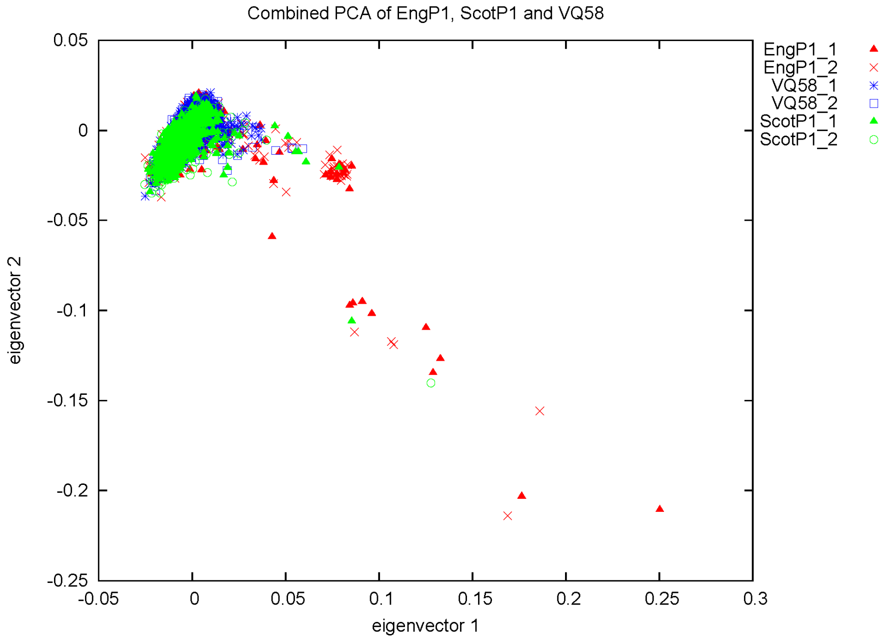


3.3.1.2 PCA of EngP1, ScotP1 and VQ58

The level of detectable population structure depends on the number and origin of samples included in the analysis. Therefore, although no structure was discovered in each individual dataset, I performed a PCA of the three main GWA study datasets, EngP1, ScotP1 and VQ58, to ensure that no structure existed within this dataset. The joint PCA revealed the existence of a group of samples (n=94) that were separated from the main cluster (see Figure 3.3). The majority of these samples, consisting of similar numbers of cases and controls, (n=62, 35 cases and 27 controls) belonged to the EngP1 dataset. The increase in the number of samples of North European descent in the analysis enhanced the differences between samples within and between datasets facilitating the detection of structure.

Figure 3.3 PC1 and PC2 plot for the combined EngP1, ScotP1 and VQ58 datasets

Cases and controls are plotted separately (1=control, 2=case). The cluster of samples separate from the main group belong largely to EngP1 and are of Jewish and Greek descent.



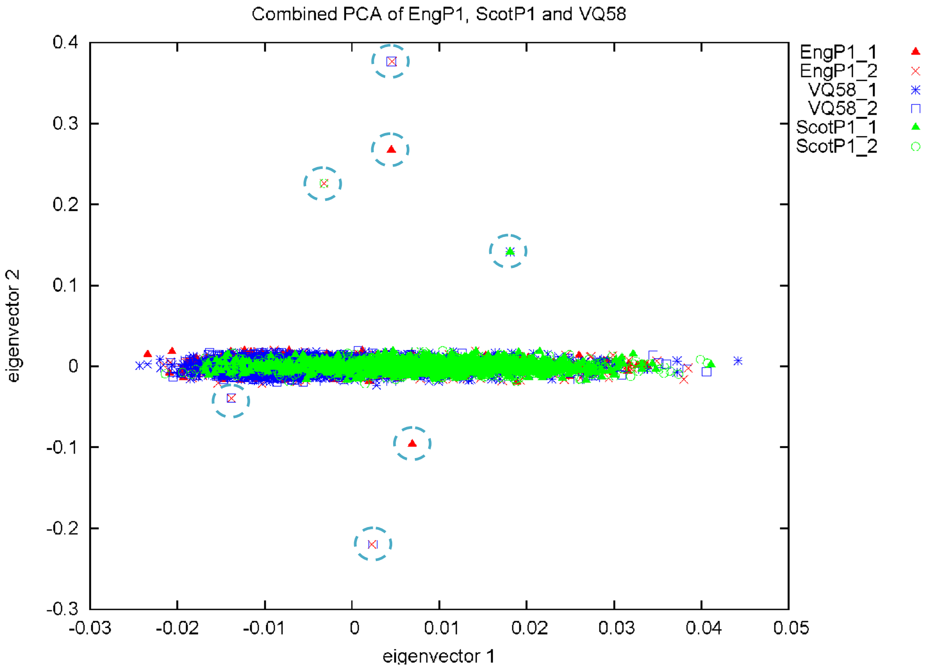
Further investigation of the ethnicity of the samples that made up the main outlying cluster revealed that most of these samples were likely from Jewish or Greek descent. The identified outliers were excluded from all subsequent analyses. A further PCA of EngP1, after the removal of these outliers, also identified two pairs of related individuals that went previously undetected. The cases were confirmed to be sisters and the controls are both from the same region of the UK, but their relationship was not confirmed.

3.3.1.3 The identification of duplicate samples between datasets during PCA

Once the initial outliers from the PCA of all three GWA datasets were removed, further interesting features were revealed. As these datasets have been recruited from the same country, there was always the possibility that there might be some overlap between the samples. The PCA highlighted seven samples as outliers to the main cluster and are circled in Figure 3.4. Each are duplicates, most of which are present in two different datasets and, therefore, would not have been identified without analysing all three datasets together. The duplicate samples were double checked using dates of births, which were identical for each pair of samples. This resulted in the removal of seven samples, one sample from each duplicate pair.

Figure 3.4 PCA plot of EngP1, ScotP1 and VQ58 showing duplicate samples across datasets

Cases and controls are plotted separately (1=control, 2=case). The circled points relate to the seven duplicate samples that were identified from this analysis, most from different datasets.



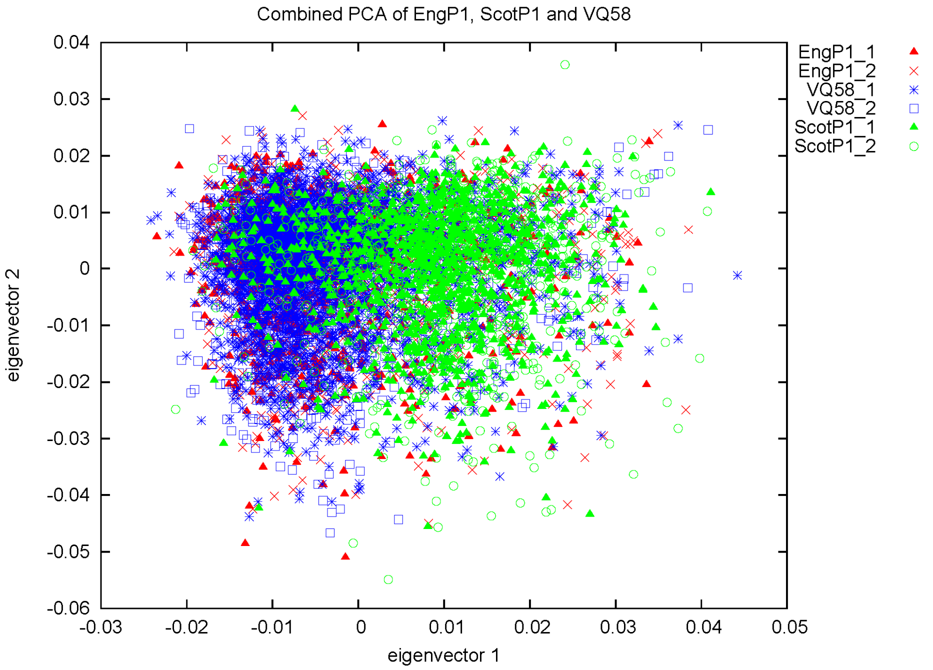
3.3.1.4 Highlighting variation between the Scottish and English datasets

Once all of the outliers described above were removed from the analysis, the plot began to highlight the subtle differences between the Scottish (ScotP1) and mainly English (EngP1 and VQ58) samples.

Figure 3.5 illustrates that the ScotP1 samples overlap with the other datasets forming a large cluster, but most are a subset of the overall population. This variation in the first principal component reflects the subtle differences in allele frequency from the North to the South of the UK.

Figure 3.5 PCA plot showing the EngP1, ScotP1 and VQ58 datasets with all major outliers removed

The plot is shown with a much smaller scale than the previous plots to gain resolution and emphasise the variation. The mean for EV1 is 2.922x10^{-7} and the standard deviation (SD) is 0.0113. The cluster spans from 2.2 SDs on the left side of the plot to 3.92 SDs away from the mean on the right side. Cases and controls are plotted separately (1=control, 2=case).



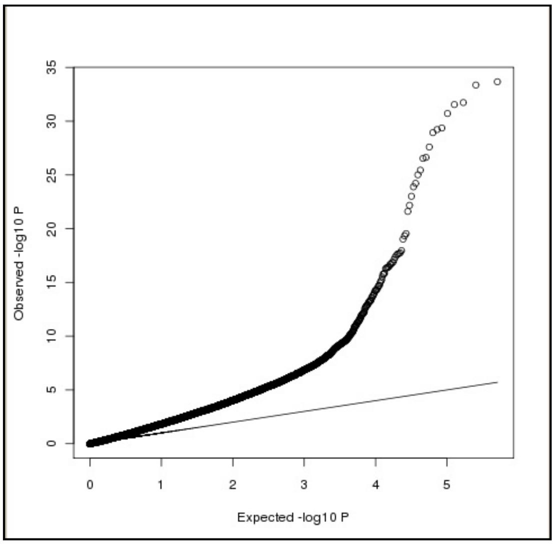
These results demonstrate the need to combine association result from the different datasets by meta-analysis in order to take into account the slight differences between datasets rather than simply combining the genotype counts. As a result of these analyses a number of individuals were removed from the GWA study and the final numbers for each dataset are given in Table 3.1 above. The alternative option to removing the samples from the analysis would have been to correct for the difference between samples using the eigenvalues as covariates in the GWA analysis. However, this could result in a smoothing of the data that may remove variation between cases and controls that is caused by disease status.

3.3.1.5 Population stratification in the Australian dataset

We attempted to include an additional dataset of cases and controls from an Australian study of CRC. All samples were from North European descent, but the controls were largely from Melbourne and the cases were slightly more diverse although still from the same region. Before incorporating these samples into our large GWAS, the samples underwent standard QC analyses. However, a QQ plot of the association statistics after the removal of SNPs out of HWE showed that there was a marked inflation, lambda=2.2, of the test statistics (see Figure 3.6) indicating population stratification.

Figure 3.6 Australian dataset QQ plot

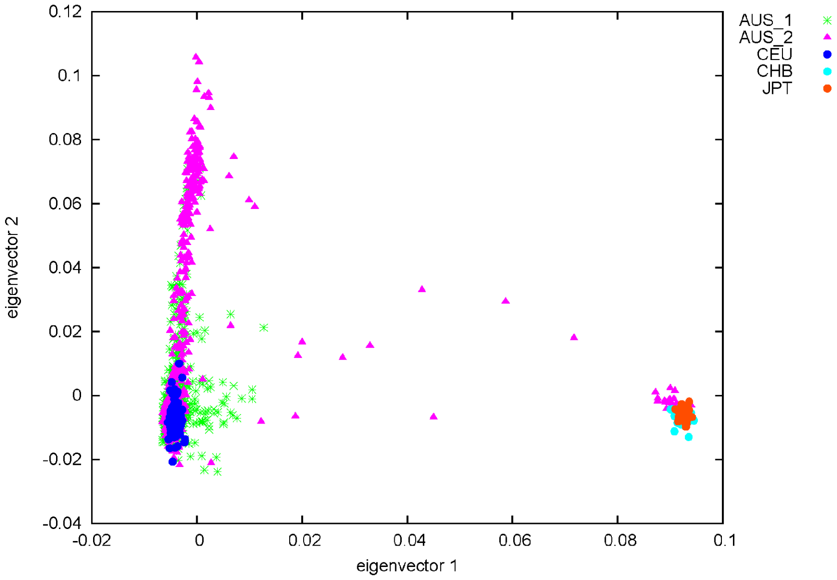
This figure only includes data from SNPs that passed the standard quality control criteria, including HWE and shows clear deviation from the expected distribution, showing marked inflation of test statistics, λ=2.2, which indicates population stratification.



The PCA of the Australian GWAS data with the addition of the HapMap CEU, CHB and JPT samples identified 28 cases that were of CHB or JPT origin (see Figure 3.7). In order to improve the resolution, the YRI HapMap population were removed from the plot.

Figure 3.7 PCA showing the Australia dataset and the HapMap CEU, CHB and JPT

The CEU, CHB and JPT samples were added to this analysis to show whether groups of samples cluster with another known population. 28 cases were of CHB or JPT origin, but there is also evidence of stratification between cases and controls along the eigenvector2 axis. Cases and controls are plotted separately (1=control, 2=case).

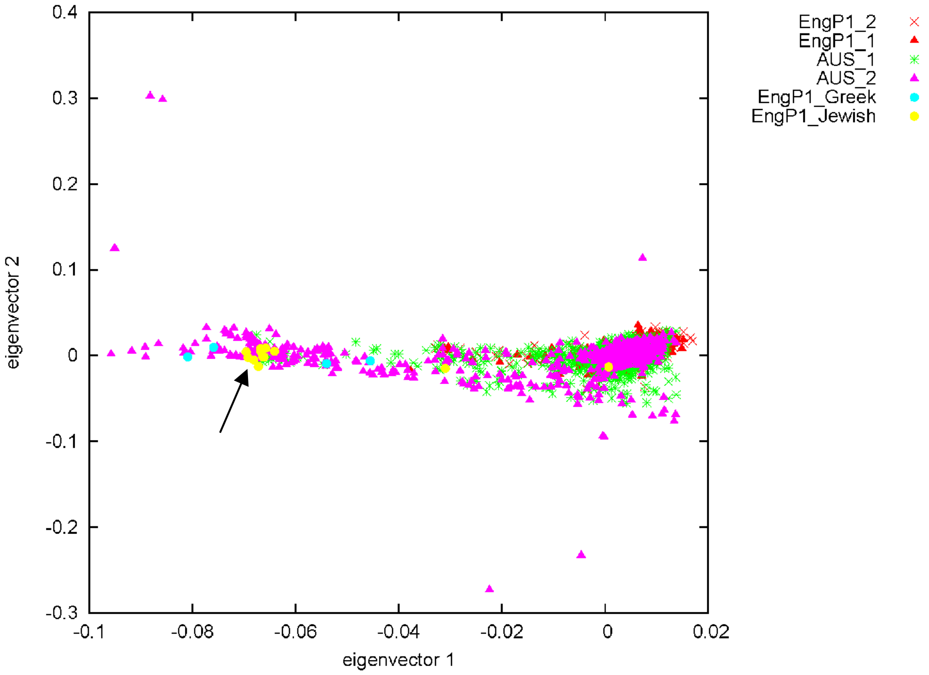


However, the removal of these outliers did not improve the level of inflation. There were 250 samples with EV2 greater than 0.02 that appear to be descended from a non-European population; there are no CEU samples beyond this point. These include 178 affected Australia samples (29% of cases), which explains the population stratification.

A further analysis, after the removal of the 28 outliers and the inclusion of EngP1 to the analysis, reveals that the majority of the Australian controls form a cluster with the EngP1 dataset to the right of the plot, while the cases are spread across the horizontal axis (see Figure 3.8). Several Greek and Jewish samples were added to the plot and can be seen clustered with Australian cases, plotted as the yellow and turquoise points at EV1 -0.07. As a result of the poor matching of cases and controls, this dataset was not included in the study.

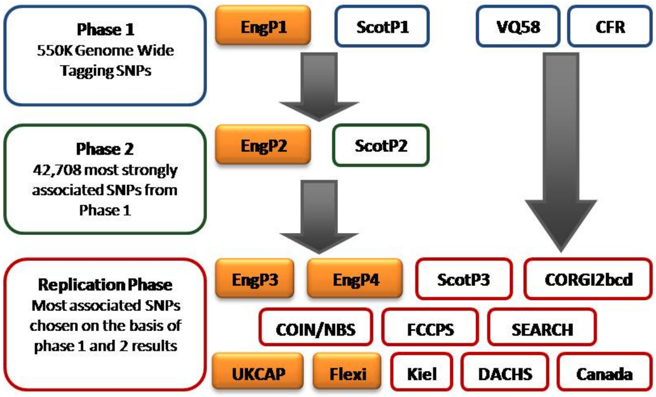
Figure 3.8 The combined PCA with the Australian (AUS) dataset and EngP1

Jewish (EngP1_Jewish), indicated by the arrow, and Greek (EngP1_Greek) samples were added to the analysis to determine if samples from these populations were included to the Australian cases leading to stratification.



3.4 Initial GWA study results

Figure 3.9 Datasets included in the initial GWA analysis


Phase 1
550K Genome Wide Tagging SNPs
EngP1
ScotP1
VQ58
CFR
Phase 2
42,708 most strongly associated SNPs from Phase 1
EngP2
ScotP2
Replication Phase
Most associated SNPs chosen on the basis of phase 1 and 2 results
EngP3
EngP4
ScotP3
CORGI2bcd
COIN/NBS
FCCPS
SEARCH
UKCAP
Flexi
Kiel
DACHS
Canada

The sample datasets used in this analysis are indicated in Figure 3.9. The strongest associations determined from the allelic P values in EngP1 were rs6983267, at chromosome 8q24 (Tomlinson *et al.* 2007), and rs4939827 at chromosome 18q21 (Broderick *et al.* 2007). These two SNPs were fast tracked to the replication phase, before the genotyping of the Phase 2 SNPs in EngP2 was complete, where the detected associations were confirmed (see Table 3.2 and Table 3.3). The replication and fine mapping of these two SNPs was divided between us (8q24) and our collaborator (18q21).

The SNP rs6983267 is not located within a gene and I was involved in the fine-mapping of the 8q24 region for which 17 additional SNPs, not on the original array, were genotyped in the EngP1 samples. I genotyped the SNP rs10505477 and the results gave an allelic P value of 7.6×10^{-6} (OR=1.32, 95% CI 1.18-1.53). In this dataset, rs6983267 has an allelic P value of 1.86×10^{-7} (OR=1.41, 95% CI 1.237-1.598). The SNP, rs10505477 is in high LD with rs6983267 (r^2=0.92, in EngP1 samples) and was the only other SNP showing a significant association with disease (see Figure 3.10). However, as these SNPs are in high pair-wise LD, it is possible that both SNPs are tagging the same causal variant. The addition of rs10505477 to a logistic regression analysis with rs6983267 significantly improved the fit of the model (P=5.22×10^{-4}, OR=1.15).

For this analysis, two datasets were added to the replication phase that included only cases that were affected with high risk adenomas (UKCAP and Flexi). A number of the cases in EngP1 had a family history of CRC, but were affected with high risk adenomas. These samples were jointly analysed in a meta-analysis to determine whether CRC risk by rs6983267 was caused by increased susceptibility to adenoma development. The results of this analysis support this hypothesis (P=6.89×10^{-5}, OR=1.22, see Table 3.2).

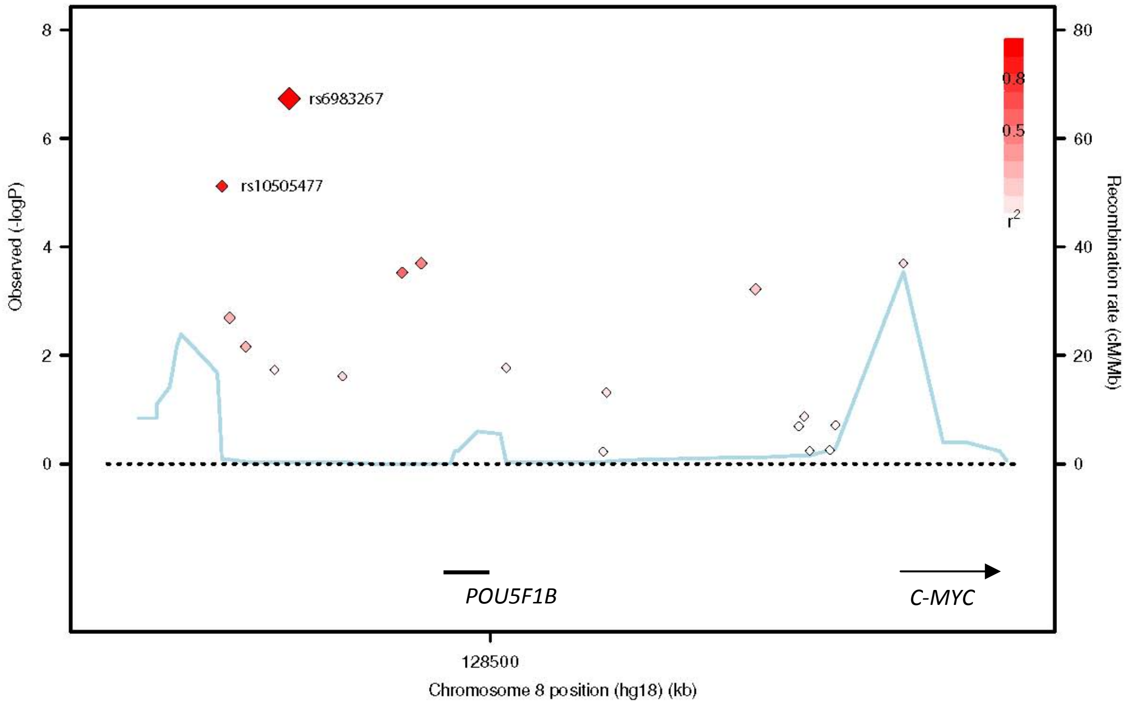
Table 3.2 The GWA results for the chromosome 8 SNPs

The ORs in this table are given with reference to the risk (major) allele. Alleles are given as Major/Minor). The UKCAP and Flexi datasets consist of samples affected with adenomas only and were, thus, not included in the combined CRC analysis. Combined analysis P values were generated using a fixed effects model. The P value for between study heterogeneity was 0.3.

SNP	Position (bp)	Alleles	Group	P value (allelic)	OR	95% CI	MAF Cases	MAF Ctrls
rs6983267 8q24	128,482,487	G/T	EngP1 (all)	1.86×10^{-7}	1.43	1.26-1.63	0.421	0.510
			EngP1 (adenomas)	5.7×10^{-7}	1.53	1.29-1.81	0.405	0.510
			EngP2	5.02×10^{-8}	1.19	1.12-1.26	0.440	0.483
			EngP3	3.42×10^{-4}	1.21	1.09-1.35	0.428	0.476
			EngP4	0.15	1.13	0.96-1.33	0.450	0.480
			UKCAP	0.51	1.05	0.90-1.23	0.455	0.491
			Flexi	0.21	1.13	0.93-1.37	0.463	0.467
			Combined CRC Analysis	1.27×10^{-14}	1.21	1.15-1.27		
			Combined Adenoma Analysis	**6.89×10^{-5}**	**1.22**	**1.10-1.34**		

Figure 3.10 Association results for the Chromosome 8 fine-mapping SNPs in EngP1

The –log P values for the chr8 region showing LD (r^2) between SNPs in relation to the rs6983267 SNP.

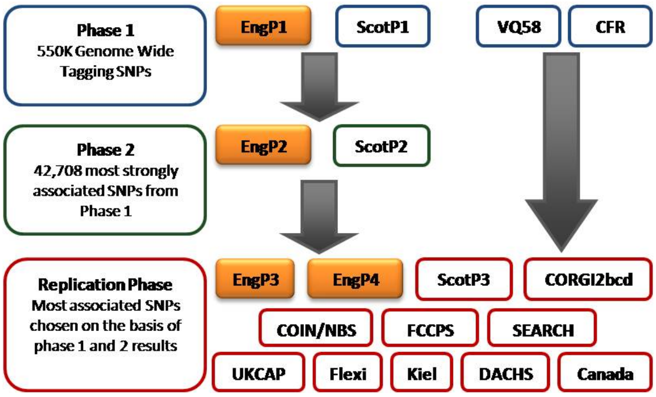


The 8q24 region does not contain any characterised genes and is referred to as a 'gene desert'. However, in addition to influencing the risk of CRC, this region was also identified in relation to increased risk of prostate (Amundadottir *et al.* 2006; Haiman *et al.* 2007; Yeager *et al.* 2007), breast (Easton *et al.* 2007), ovarian (Ghoussaini *et al.* 2008) and urinary bladder cancer (Kiemeney *et al.* 2008). The SNP rs6983267 is also significantly associated with an increased risk of prostate cancer (Haiman *et al.* 2007) and ovarian cancer (Ghoussaini *et al.* 2008). Analysis of the LD pattern surrounding rs6983267 reveals that it tags a processed pseudogene of the *OCT4* transcription factor, POU5F1, although no associated SNP were identified within this gene. This 'pseudogene' has recently been reclassified as POU5F1B as it transcribes a functional protein that acts as a weak transcriptional activator with strong similarity to POU5F1 (Panagopoulos *et al.* 2008) and is over-expressed in prostate cancers (Kastler *et al.* 2010).

The nearest characterised gene, proto-oncogene *MYC,* is located 116kb telomeric to rs6983267, but there was no significant association between SNPs mapping to the *MYC* and CRC. However, recent work on the specific regions identified in each of the cancer types above has shown that the loci are located within gene regulatory elements that bear epigenetic chromatin marks of enhancer elements (Sotelo *et al.* 2010). Another group tested the physical interaction of these elements with *MYC* and demonstrated that these loci are regulatory elements that exhibit long range enhancer effects on *MYC* in a tissue specific manner (Ahmadiyeh *et al.* 2010).

3.4.1 The chromosome 18 locus

Figure 3.11 Datasets included in the chr15 and chr18 analyses


Phase 1
550K Genome Wide Tagging SNPs
EngP1
ScotP1
VQ58
CFR
Phase 2
42,708 most strongly associated SNPs from Phase 1
EngP2
ScotP2
Replication Phase
Most associated SNPs chosen on the basis of phase 1 and 2 results
EngP3
EngP4
ScotP3
CORGI2bcd
COIN/NBS
FCCPS
SEARCH
UKCAP
Flexi
Kiel
DACHS
Canada

The second most strongly associated SNP that was fast tracked to the replication phase was located at 44,707,461bp on chromosome 18 within the gene Mothers against decaplentaplegic 7 (*SMAD7*). The datasets included in this analysis are given in the figure above. The replication phase analysis results (given in Table 3.3) confirmed the association of rs4939827 with CRC (P=1.00x10^{-12}, OR=0.85).

Table 3.3 The GWA results for the chromosome 18 SNPs

A summary of the association results for the most strongly associated SNPs. The combined replication P value is for the joint analysis of Eng1 and the replication phase. The SNP positions are from genome build 35. Alleles are given as minor/major and the OR is calculated with reference to the minor allele. The combined analysis only included cancer cases and the P value was generated using a fixed effects model.

SNP	Chr.	Position (bp)	Alleles	Group	P (trend)	OR	95% CI	MAF Cases	MAF Ctrls
rs4939827	18q21	44,707,461	T/C	EngP1	3.07x10^{-7}	0.71		0.406	0.489
				EngP2	1.42x10^{-4}	0.89		0.439	0.469
				EngP3	7.72x10^{-6}	0.81		0.441	0.494
				EngP4	0.280	0.91		0.449	0.471
				Combined	1.00x10^{-12}	0.85	0.81-0.89		
rs12953717	18q21	44,707,927	C/T	EngP1	1.07x10^{-6}	1.38		0.496	0.417
				EngP2	2.69x10^{-6}	1.16		0.469	0.432
				EngP3	6.74x10^{-3}	1.14		0.460	0.428
				EngP4	0.481	1.06		0.458	0.443
				Combined	9.10x10^{-12}	1.17	1.12-1.22		

3.4.2 The identification of chromosome 15 SNPs in the *HMPS* locus

There were two additional SNPs that were also fast tracked to the replication phase owing to the SNPs close proximity to the previously identified *HMPS*/*CRAC1* locus located on chromosome 15q13, which was found through linkage analysis to be strongly associated with CRC in the Ashkenazi Jewish population (Jaeger *et al.* 2003).

The SNPs rs4779584 and rs10318 displayed a modest association in EngP1 (P=1.31x10^{-3} and P=0.01, respectively). However, this association was confirmed in the replication phase (including the datasets shown in Figure 3.11) after achieving allelic P values of 4.44x10^{-14} for rs4779584 and 7.93x10^{-9} for rs10318 (see Table 3.4)(Jaeger *et al.* 2008). These SNPs are in moderate LD (r^{2}=0.57, D'=0.77) with no evidence of independent effects after a logistic regression analysis showed that the inclusion of rs10318 did not improve the model compared to rs4779584 alone (P=0.47).

Table 3.4 The GWA results for the chromosome 15 SNPs

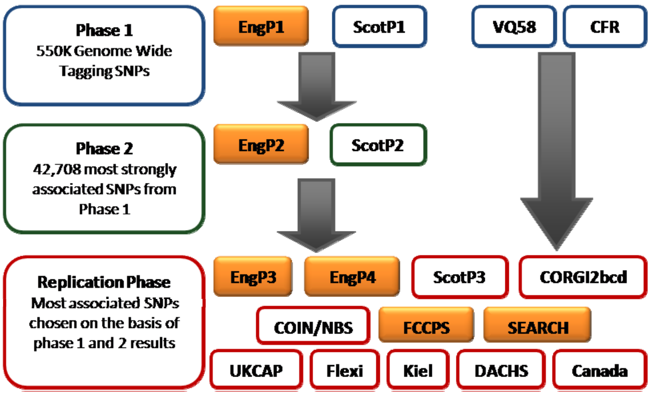
SNP positions are from the genome build 35 and ORs were calculated with reference to the minor allele. EngP1 in this analysis only included cases with CRC and was supplemented by cases from the VICTOR trial, which are part of the VQ58 dataset. EngP4 consists of CORGI2bc and additional VICTOR cases. Combined P values generated using a fixed effects model.

SNP	Chr.	Position (bp)	Alleles	Group	P (allelic)	OR	95% CI	MAF Cases	MAF Ctrls
rs4779584	15q13	30,782,048	T/C	EngP1+V	4.34x10^{-4}	1.35	1.14-1.60	0.234	0.184
				EngP2	4.91x10^{-7}	1.21	1.13-1.31	0.222	0.190
				EngP3	7.05x10^{-8}	1.39	1.23-1.57	0.222	0.170
				EngP4	0.439	1.09	0.87-1.38	0.204	0.109
				Combined	**4.44x10^{-14}**	**1.26**	**1.19-1.34**		
rs10318	15q13	30,813,271	A/G	EngP1+V	6.97x10^{-3}	1.26	1.06-1.51	0.219	0.182
				EngP2	3.66x10^{-5}	1.18	1.09-1.27	0.210	0.184
				EngP3	9.01x10^{-4}	1.22	1.08-1.37	0.218	0.187
				EngP4	0.729	1.04	0.82-1.32	0.187	0.181
				Combined	**7.93X10^{-9}**	**1.19**	**1.12-1.26**		

The SNP rs10318 is located in the 3' untranslated region of gremlin-1 precursor (*GREM1*, also known as *DRM*), which encodes a secreted bone morphogenic protein (BMP) antagonist. The SNP rs4779584 is located between the genes Secretogranin V isoform 2 (*SCG5,* also known as SGNE1) and *GREM1*. All of these genes have plausible functional effects to implicate them in CRC risk. These three genes were divided into fragments and screened for variations using 96 CRC cases from EngP1 (half of which were diagnosed under the age of 40) and a combination of the LightScanner (Idaho Technology Inc., described in Section 2.9) and standard sequencing. This work was divided between the first authors of the paper. However, although a number of novel variations were identified, the causal mutation(s) is yet to be determined.

3.5 The combined analysis of EngP1 and EngP2

Figure 3.12 Datasets included in the analysis of EngP1 and EngP2


Phase 1
550K Genome Wide Tagging SNPs
EngP1
ScotP1
VQ58
CFR
Phase 2
42,708 most strongly associated SNPs from Phase 1
EngP2
ScotP2
Replication Phase
Most associated SNPs chosen on the basis of phase 1 and 2 results
EngP3
EngP4
ScotP3
CORGI2bcd
COIN/NBS
FCCPS
SEARCH
UKCAP
Flexi
Kiel
DACHS
Canada

The SNPs selected for genotyping in the EngP2 samples were jointly analysed with the EngP1 results using the Mantel-Haenszel method for meta-analysis (Pettiti 1994). A further eleven SNPs were identified with a P value less than 1x10^{-4} and were followed up in EngP3. These SNPs were rs16892766 (8q23.3), rs10795668 (10p14), rs4355419 (4q13.1), rs2488704 (10q22.1), rs2282428 (1q42.2), rs12957142 (18q12.3), rs4822442

(22q11.23), rs11590577 (1p36.31), rs4841306 (8q23.10), rs2164182 (11q21) and rs2989734 (9q34.30). However, just two SNPs, rs10795668 and rs16892766, had their association successfully confirmed in the replication phase achieving overall P values of 2.5x10^{-13} and 9.6x10^{-18}, respectively (see Table 3.5)(Tomlinson *et al.* 2008). For this study the replication phase also included ScotP1, ScotP2, EPICOLON, and the additional datasets DFCCS, MCCS, POPGENSHIP and EPICOLON (see section 2.1 and the appendix).

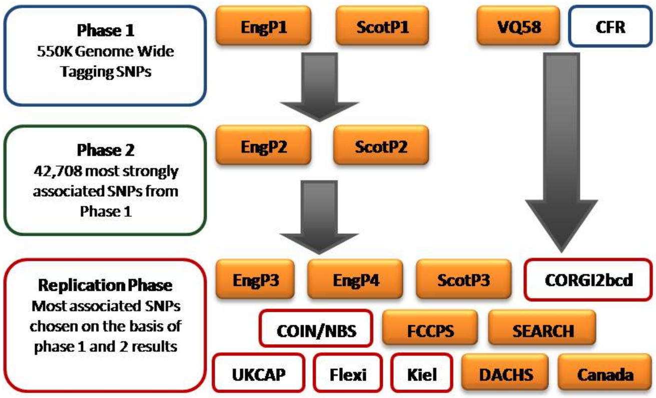
Table 3.5 Summary of the two associated SNPs from the analysis of EngP1 and EngP2

The summary statistics for the two most strongly associated SNPs identified in the meta-analysis of EngP1 and EngP2. The alleles are given as minor/major. The combined P values were generated by meta-analysis using the fixed effects model.

SNP	Position (bp)	Alleles	Group	P (allelic)	OR	95% CI	MAF Cases	MAF Ctrls
rs16892766 8q23.3	117,699,995	C/A	EngP1	7.57x10^{-3}	1.37	1.09- 1.72	0.10	0.08
			EngP2	1.87x10^{-6}	1.38	1.21- 1.57	0.10	0.07
			EngP3	0.029	1.16	1.01-1.32	0.089	0.078
			EngP4	0.186	1.17	0.93-1.47	0.089	0.077
			SEARCH	7.76x10^{-5}	1.36	1.17-1.58	0.093	0.070
			EPICOLON	0.519	0.90	0.65-1.25	0.072	0.079
			FCCPS	0.041	1.21	1.01-1.46	0.141	0.119
			POPGENSHIP	1.09x10^{-3}	1.26	1.10-1.15	0.096	0.077
			DFCCS	0.022	1.35	1.04-1.76	0.106	0.081
			MCCS	0.587	1.08	0.81-1.45	0.088	0.081
			ScotP1	0.067	1.22	0.99-1.51	0.102	0.085
			ScotP2	8.24x10^{-4}	1.29	1.11-1.50	0.104	0.082
EngP1/EngP2 combined				1.93x10^{-8}	1.41	1.25-1.56		
All Combined				**3.3x10^{-18}**	**1.25**	**1.19-1.32**		
rs10795668 10p14	8,741,225	A/G	EngP1	6.06x10^{-3}	0.82	0.71- 0.95	0.30	0.34
			EngP2	5.57x10^{-3}	0.89	0.83- 0.97	0.30	0.33
			EngP3	6.94x10^{-6}	0.84	0.77-0.90	0.290	0.328
			EngP4	0.061	0.88	0.77-1.01	0.300	0.328
			SEARCH	8.54x10^{-5}	0.84	0.77-0.91	0.300	0.338
			EPICOLON	0.160	0.87	0.72-1.06	0.280	0.309
			FCCPS	0.0138	0.84	0.73-0.96	0.271	0.307
			POPGENSHIP	0.296	0.96	0.88-1.04	0.323	0.333
			ScotP1	0.0793	0.89	0.77-1.01	0.298	0.324
			ScotP2	0.758	0.99	0.90-1.08	0.325	0.328
EngP1/EngP2 combined				6.99x10^{-5}	0.72	0.78-0.86		
All Combined				**2.5x10^{-13}**	**0.89**	**0.86-0.91**		

3.6 The meta-analysis of the English and Scottish GWA studies

Figure 3.13 The datasets included in the meta-analysis of the English and Scottish GWA studies


Phase 1
550K Genome Wide Tagging SNPs
EngP1
ScotP1
VQ58
CFR
Phase 2
42,708 most strongly associated SNPs from Phase 1
EngP2
ScotP2
Replication Phase
Most associated SNPs chosen on the basis of phase 1 and 2 results
EngP3
EngP4
ScotP3
CORGI2bcd
COIN/NBS
FCCPS
SEARCH
UKCAP
Flexi
Kiel
DACHS
Canada

In order to increase the power to detect variants with smaller effect than those already identified (and in concert with our collaborators at the ICR in Sutton), I performed a meta-analysis to combine the EngP1 and EngP2 results with the results from the two phases of our collaborator's GWA study, ScotP1 and ScotP2. The results identified nine SNPs for additional study. At this point it was useful to test these SNPs for association in an additional GWA dataset to inform the decision of which SNPs would be taken forward for further analysis. Therefore, I combined the VQ58 results in a meta-analysis with the four datasets above. The analysis resulted in all nine SNPs being taken forward for additional genotyping in the replication phase, which was divided between the groups contributing to the work (as part of this work, I genotyped three SNPs, rs961253, rs9929218, rs1862748 in the CORGI2bc samples).

The results for the individual datasets were combined by meta-analysis with the replication phases, including VQ58, by our collaborators at the ICR. All SNPs, except those on chromosome 1, were successfully replicated (Houlston *et al.* 2008)(see Table 3.6). The results of this analysis confirmed the association of the most significant SNPs identified in the initial GWA analysis, but also highlighted additional regions of association that merit further study. Prominent among these from a functional point of view are *E-cadherin* (*CDH1*) and *P-cadherin* (*CDH3*) on chromosome 16 and *BMP4* on chromosome 14 for its proximity to the CRAC1/HMPS locus. The nearest gene to the SNPs on chromosome 20 is bone morphogenetic protein two preprotein (*BMP2)*, which belongs to the TGFβ superfamily, and those on chromosome 19 are located within Rhophilin (*RHPN2*), which is a Rho GTPase binding protein.

Table 3.6 The meta-analysis results for the SNPs in the five susceptibility loci

The alleles in the table are given as minor/major alleles and the odds ratios are in relation to the minor allele. The combined phases included the replication datasets and EngP1. Combined P values were generated using the fixed effects model.

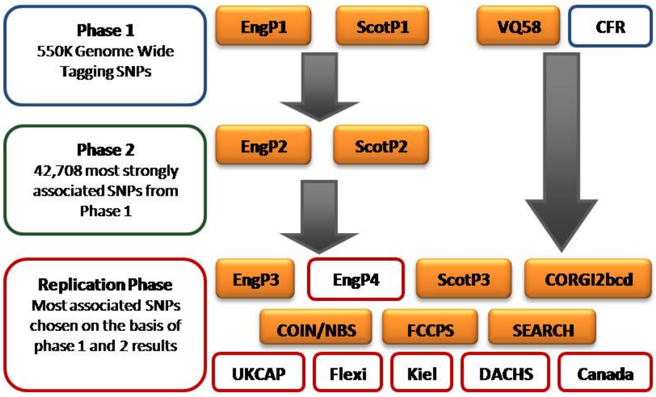
Chr.	SNP	Position (bp)	Alleles	Group	P (allelic)	OR	95% CI	MAF Cases	MAF Ctrls
20	rs961253	6,352,281	A/C	EngP1	0.13	1.12	0.97-1.30	0.397	0.369
				ScotP1	5.7×10^{-2}	1.13	0.99-1.29	0.390	0.361
				EngP2	5.6×10^{-3}	1.11	1.03-1.20	0.381	0.355
				ScotP2	8.6×10^{-4}	1.17	1.06-1.27	0.383	0.347
				EngP3	3.2×10^{-3}	1.12	1.04-1.20	0.376	0.350
				ScotP3	9.1×10^{-2}	1.14	0.98-1.33	0.370	0.340
				FCCPS	5.1×10^{-3}	1.23	1.07-1.42	0.339	0.294
				VCQ58	0.386	1.04	0.95-1.15	0.378	0.368
			P1/P2 Combined		8.9×10^{-7}	1.13	1.08-1.19		
			All Combined		**2.0×10^{-10}**	**1.12**	**1.08-1.16**		
20	rs355527	6,336,068	A/G	EngP1	0.20	1.10	0.95-1.23	0.357	0.335
				ScotP1	6.0×10^{-2}	1.13	0.99-1.29	0.353	0.324
				EngP2	1.4×10^{-3}	1.14	1.05-1.23	0.360	0.332
				ScotP2	3.8×10^{-3}	1.14	1.04-1.25	0.356	0.326
				EngP3	4.9×10^{-3}	1.11	1.03-1.20	0.345	0.322
				ScotP3	0.12	1.13	0.97-1.32	0.345	0.319
				FCCPS	1.3×10^{-2}	1.21	1.04-1.40	0.322	0.282
				VCQ58	0.158	1.07	0.97-1.18	0.353	0.337
			P1/P2 Combined		1.2×10^{-6}	1.13	1.08-1.19		
			All Combined		**2.1×10^{-10}**	**1.12**	**1.08-1.17**		
14	rs4444235	53,480,669	C/T	EngP1	1.2×10^{-3}	1.27	1.10-1.47	0.511	0.452

				ScotP1	5.4x10^{-2}	1.13	0.99-1.28	0.481	0.451
				EngP2	6.9x10^{-3}	1.11	1.03-1.19	0.487	0.461
				ScotP2	3.3x10^{-3}	1.10	1.00-1.20	0.478	0.454
				EngP3	4.2x10^{-3}	1.11	1.03-1.19	0.491	0.466
				ScotP3	0.10	1.13	0.98-1.30	0.488	0.458
				FCCPS	3.1x10^{-2}	1.16	1.02-1.33	0.462	0.426
				VCQ58	0.286	1.05	0.96-1.15	0.479	0.466
				P1/P2 Combined	1.8x10^{-6}	1.12	1.07-1.18		
				All Combined	**8.1x10^{-10}**	**1.11**	**1.08-1.15**		
19	rs10411210	38,224,140	T/C	EngP1	1.6x10^{-2}	0.72	0.55-0.94	0.072	0.097
				ScotP1	1.8x10^{-4}	0.64	0.50-0.81	0.061	0.093
				EngP2	1.6x10^{-2}	0.85	0.75-0.97	0.084	0.097
				ScotP2	2.7x10^{-3}	0.79	0.67-0.92	0.076	0.095
				EngP3	0.02	0.87	0.77-0.98	0.085	0.097
				ScotP3	3.9x10^{-3}	0.66	0.50-0.88	0.062	0.090
				FCCPS	0.10	0.85	0.70-1.03	0.138	0.158
				VCQ58	0.88	0.98	0.84-1.15	0.093	0.094
				Canada	0.42	0.92	0.76-1.12	0.097	0.104
				DACHS	0.36	1.08	0.91-1.28	0.110	0.103
				Kiel	0.14	0.89	0.77-1.04	0.082	0.091
				SEARCH	0.12	0.88	0.78-1.03	0.084	0.094
				P1/P2 Combined	4.9x10^{-8}	0.79	0.72-0.86		
				All Combined	**4.6x10^{-9}**	**0.87**	**0.83-0.91**		
16	rs9929218	67,378,447	A/G	EngP1	7.5x10^{-2}	0.87	0.74-1.01	0.282	0.312
				ScotP1	1.4x10^{-2}	0.84	0.73-0.97	0.266	0.301
				EngP2	1.9x10^{-2}	0.91	0.84-0.98	0.275	0.295
				ScotP2	1.7x10^{-3}	0.86	0.78-0.94	0.263	0.294
				EngP3	4.9x10^{-2}	0.92	0.86-0.99	0.280	0.296
				ScotP3	0.37	0.93	0.79-1.09	0.257	0.272
				FCCPS	0.782	0.97	0.83-1.14	0.227	0.231
				VCQ58	0.112	0.92	0.83-1.02	0.284	0.301
				Canada	1.3x10^{-2}	0.85	0.75-0.97	0.274	0.307
				DACHS	0.61	0.97	0.86-1.09	0.286	0.293
				Kiel	0.94	0.99	0.91-1.09	0.281	0.282
				SEARCH	4.9x10^{-2}	0.91	0.83-0.99	0.281	0.300
				P1/P2 Combined	1.4x10^{-6}	0.88	0.83-0.92		
				All Combined	**1.2x10^{-8}**	**0.91**	**0.89-0.94**		
16	rs1862748	67,390,444	T/C	EngP1	8.8x10^{-2}	0.87	0.75-1.02	0.296	0.326
				ScotP1	5.0x10^{-3}	0.82	0.72-0.94	0.280	0.321
				EngP2	2.9x10^{-2}	0.91	0.84-0.99	0.290	0.308
				ScotP2	3.8x10^{-3}	0.87	0.79-0.96	0.280	0.309
				EngP3	3.4x10^{-2}	0.92	0.85-0.99	0.292	0.310
				ScotP3	0.30	0.92	0.78-1.08	0.267	0.284
				VCQ58	0.87	1.02	0.83-1.24	0.314	0.310
				Canada	8.9x10^{-2}	0.90	0.79-1.02	0.294	0.317
				DACHS	0.62	0.97	0.87-1.09	0.303	0.310
				Kiel	0.42	0.96	0.88-1.06	0.293	0.301
				SEARCH	3.2x10^{-2}	0.91	0.83-0.99	0294	0.315
				P1/P2 Combined	2.6x10^{-6}	0.88	0.84-0.93		

				All Combined	**2.9×10^{-8}**	**0.91**	**0.88-0.94**		
19	rs7259371	38,226,481	A/G	EngP1	9.2×10^{-2}	0.85	0.70-1.03	0.160	0.184
				ScotP1	1.2×10^{-3}	0.76	0.64-0.90	0.145	0.183
				EngP2	2.3×10^{-2}	0.89	0.81-0.98	0.165	0.182
				ScotP2	2.5×10^{-2}	0.88	0.78-0.98	0.160	0.178
				EngP3	7.1×10^{-3}	0.88	0.80-0.97	0.169	0.188
				ScotP3	2.2×10^{-3}	0.73	0.60-0.89	0.135	0.176
				FCCPS	0.282	0.91	0.77-1.08	0.196	0.212
				VCQ58	0.383	1.06	0.94-1.19	0.177	0.169
				P1/P2 Combined	5.7×10^{-6}	0.86	0.81-0.92		
				All Combined	**2.2×10^{-7}**	**0.89**	**0.85-0.93**		
1	rs4951291	202,273,161	T/C	EngP1	0.73	0.96	0.78-1.19	0.129	0.134
				ScotP1	7.2×10^{-2}	0.85	0.71-1.02	0.128	0.147
				EngP2	1.7×10^{-2}	0.88	0.79-0.98	0.132	0.148
				ScotP2	9.2×10^{-5}	0.78	0.69-0.88	0.126	0.157
				EngP3	0.22	1.07	0.96-1.18	0.142	0.135
				ScotP3	0.47	0.92	0.75-1.14	0.133	0.142
				FCCPS	0.84	0.98	0.82-1.18	0.149	0.151
				VCQ58	0.99	0.99	0.87-1.14	0.136	0.136
				P1/P2 Combined	5.3×10^{-6}	0.85	0.79-0.91		
				All Combined	**4.1×10^{-3}**	**0.93**	**0.88-0.98**		
1	rs4951039	202,273,220	G/A	EngP1	0.85	0.98	0.79-1.21	0.132	0.134
				ScotP1	8.2×10^{-2}	0.85	0.71-1.02	0.129	0.148
				EngP2	1.9×10^{-2}	0.88	0.79-0.98	0.134	0.150
				ScotP2	5.2×10^{-5}	0.77	0.68-0.88	0.127	0.158
				EngP3	3.7×10^{-2}	1.11	1.01-1.23	0.149	0.136
				VCQ58	0.64	1.04	0.89-1.21	0.147	0.142
				P1/P2 Combined	5.8×10^{-6}	0.85	0.79-0.91		
				All Combined	**2.4×10^{-2}**	**0.94**	**0.89-0.99**		

3.7 The meta analysis of 3 GWA studies and imputed SNPs in VQ58 to discover new SNPs

Figure 3.14 The datasets included in the 3 GWA study meta-analysis


Phase 1
550K Genome Wide Tagging SNPs
EngP1
ScotP1
VQ58
CFR
Phase 2
42,708 most strongly associated SNPs from Phase 1
EngP2
ScotP2
Replication Phase
Most associated SNPs chosen on the basis of phase 1 and 2 results
EngP3
EngP4
ScotP3
CORGI2bcd
COIN/NBS
FCCPS
SEARCH
UKCAP
Flexi
Kiel
DACHS
Canada

Up until this point, the initial phase one and phase two datasets had been used as the discovery datasets to identify new SNPs and datasets such as VQ58 were used to replicate these findings. However, as the effect sizes that have been reported in GWA studies thus far have been small, larger numbers of samples are required to achieve sufficient power to detect additional variants by this method. Therefore, to attempt to uncover new associated SNPs, the EngP1 and ScotP1 samples were combined in a meta-analysis with the VQ58 dataset. VQ58 was genotyped using the Illumina Hap300 SNP array and so the genotypes of SNPs missing from this array, but included on the Hap550 were imputed using the EngP1 controls and HapMap phase II as reference panels (this is discussed in Section 4.3, below). Overall, 401,013 SNPs were included in this analysis from those genotyped in ScotP1 and EngP1 and genotyped or imputed successfully in VQ58. A different approach was used in order to treat the imputed data appropriately (see section 4.2.3). The data was analysed initially using SNPTEST and meta-analysis of these results was undertaken using the program META, which combines the P values of the studies while taking account of the effect and sample size of each dataset (http://www.stats.ox.ac.uk/~jsliu/meta.html).

The results of a fixed-effects meta analysis of EngP1, ScotP1 and VQ58 showed that just one SNP out of those previously detected achieved genome-wide significance (defined as P less than $1x10^{-7}$), rs4939827 located on chromosome 18q21.1 ($P=3.92x10^{-9}$). In total this analysis identified 66 SNPs, shown in Table 3.7, that showed evidence of association ($P<1x10^{-4}$) with no evidence of between study heterogeneity

(P>0.05). This list contains 21 SNPs that were imputed in VQ58 and 58 SNPs that had previously been identified and genotyped in the phase two samples.

Table 3.7 The results of the meta-analysis of EngP1, ScotP1 and VQ58 with P<1x10^{-4}

This table shows the overall P value of the meta-analysis combining data, using the fixed effects model, from the three GWA studies and including SNPs that were imputed in VQ58 to achieve a better overlap of SNPs. Under the 'Type' column, genotyped SNPs are labelled 'G' and imputed SNPs are labelled 'I'. The model parameter estimates (betas) and standard errors (SE) are also given. Beta values are calculated with reference to the B allele (where the allele A is coded as 0 and allele B is coded as 1, beta is an estimate of the increase in log-odds that can be attributed to each copy of the B allele).

Chr.	SNP	Position (bp)	Allele A/B	P value	Beta (log OR)	SE	P het	Type	Genotyped in EngP2/ScotP2
18	rs4939827	44,707,461	C/T	3.92x10^{-9}	0.19	0.03	0.07	G	Y
8	rs7014346	128,493,974	A/G	5.39x10^{-7}	-0.17	0.03	0.19	G	Y
7	rs216735	28,563,719	A/G	9.36x10^{-7}	0.23	0.05	0.42	G	Y
8	rs7837328	128,492,309	A/G	1.37x10^{-6}	-0.16	0.03	0.24	I	Y
12	rs7138945	48,825,686	G/T	1.60x10^{-6}	-0.17	0.03	0.70	G	Y
15	rs4779584	30,782,048	C/T	1.91x10^{-6}	0.20	0.04	0.26	G	Y
11	rs11236164	73,972,614	A/C	3.22x10^{-6}	-0.15	0.03	0.84	I	N
12	rs11169282	48,816,238	A/G	3.25x10^{-6}	0.16	0.03	0.57	G	N
18	rs7228236	32,249,233	A/G	3.97x10^{-6}	0.19	0.04	0.41	G	Y
15	rs7182555	95,177,768	C/T	4.68x10^{-6}	-0.28	0.06	0.07	G	Y
18	rs4464148	44,713,030	C/T	6.32x10^{-6}	-0.16	0.04	0.07	G	Y
10	rs10829813	132,513,083	A/G	6.96x10^{-6}	-0.43	0.10	0.13	G	N
11	rs3824999	74,023,198	G/T	9.46x10^{-6}	-0.15	0.03	0.83	G	Y
9	rs2185857	109,343,284	A/G	1.03x10^{-5}	0.17	0.04	0.81	G	Y
4	rs16837803	5,971,290	C/T	1.12x10^{-5}	0.29	0.07	0.73	I	Y
10	rs2789310	19,368,441	A/C	1.18x10^{-5}	-0.25	0.06	0.84	G	Y
2	rs6434983	199,284,390	C/T	1.19x10^{-5}	-0.17	0.04	0.96	I	Y
2	rs1400976	199,275,159	C/T	1.20x10^{-5}	-0.17	0.04	0.97	I	Y
11	rs7128034	41,763,857	C/T	1.22x10^{-5}	0.26	0.06	0.69	G	Y
11	rs10219203	74,002,571	C/T	1.25x10^{-5}	0.15	0.03	0.77	I	Y
12	rs10492081	46,761,120	A/G	1.34x10^{-5}	-0.18	0.04	0.19	G	Y
12	rs11169552	49,441,930	C/T	1.49x10^{-5}	-0.16	0.04	0.78	G	Y
10	rs7898455	8,778,914	G/T	1.67x10^{-5}	-0.15	0.04	0.60	I	Y
18	rs1316447	44,726,674	A/G	1.73x10^{-5}	-0.19	0.04	0.93	G	Y
4	rs1381626	88,771,808	A/G	1.96x10^{-5}	-0.20	0.05	0.20	G	Y
2	rs1356494	199,281,141	A/G	2.13x10^{-5}	0.16	0.04	0.99	G	Y
7	rs13233942	28,607,958	A/G	2.18x10^{-5}	-0.17	0.04	0.89	I	Y
18	rs2000662	32,220,133	C/T	2.21x10^{-5}	-0.17	0.04	0.51	G	Y

13	rs6491545	99,590,735	C/T	2.46x10^{-5}	-0.14	0.03	0.66	I	Y
12	rs7312252	49,030,438	C/T	2.53x10^{-5}	0.15	0.03	0.73	G	Y
12	rs12303082	49,040,830	G/T	2.59x10^{-5}	0.15	0.03	0.75	G	Y
12	rs1362983	48,900,974	A/G	2.71x10^{-5}	0.15	0.03	0.74	G	Y
12	rs7134595	49,016,725	C/T	2.92x10^{-5}	0.15	0.03	0.74	I	Y
12	rs1344958	49,483,984	C/T	3.18x10^{-5}	-0.16	0.04	0.79	G	Y
10	rs706771	8,736,452	A/G	3.20x10^{-5}	0.15	0.04	0.69	G	Y
12	rs11833608	49,043,895	A/G	3.28x10^{-5}	-0.15	0.03	0.76	G	Y
11	rs1292504	73,947,278	A/G	3.28x10^{-5}	-0.15	0.04	0.55	G	Y
11	rs11236203	74,055,648	G/T	3.48x10^{-5}	-0.14	0.03	0.67	I	Y
12	rs12828340	48,923,562	C/T	3.57x10^{-5}	-0.14	0.03	0.74	I	N
12	rs6580742	49,014,078	C/T	3.75x10^{-5}	0.17	0.04	0.25	G	Y
12	rs11169335	48,922,631	A/G	3.78x10^{-5}	0.14	0.03	0.73	I	Y
7	rs11981075	83,343,328	A/G	3.82x10^{-5}	0.15	0.04	0.05	G	Y
8	rs10808555	128,478,693	A/G	3.89x10^{-5}	0.14	0.03	0.56	G	Y
3	rs10936599	170,974,795	C/T	4.05x10^{-5}	-0.16	0.04	0.23	G	Y
6	rs4945754	107,022,523	A/G	4.70x10^{-5}	-0.29	0.07	0.17	G	N
5	rs6897885	15,704,836	C/T	4.81x10^{-5}	0.19	0.05	0.39	I	Y
15	rs2053423	64,813,271	C/T	5.10x10^{-5}	0.15	0.04	0.37	G	Y
9	rs3893493	135,683,727	A/G	5.55x10^{-5}	0.15	0.04	0.83	G	Y
1	rs12037907	67,745,023	A/C	5.67x10^{-5}	0.22	0.05	0.12	I	Y
2	rs6434981	199,282,102	A/C	5.86x10^{-5}	0.13	0.03	0.11	I	Y
3	rs3772190	170,983,181	A/G	6.27x10^{-5}	0.16	0.04	0.24	I	Y
2	rs359700	126,838,951	A/G	6.33x10^{-5}	-0.22	0.05	0.08	G	N
15	rs10318	30,813,271	C/T	6.49x10^{-5}	0.16	0.04	0.10	G	Y
16	rs7404339	64,973,293	A/G	6.57x10^{-5}	-0.13	0.03	0.11	G	Y
4	rs6828852	5,967,739	C/T	6.65x10^{-5}	-0.42	0.10	0.95	I	Y
2	rs2123693	108,221,600	C/T	7.25x10^{-5}	-0.13	0.03	0.45	G	Y
2	rs360377	126,836,315	A/G	7.62x10^{-5}	0.21	0.05	0.08	G	N
2	rs1878665	199,288,971	G/T	8.62x10^{-5}	0.14	0.04	0.99	G	Y
7	rs2708594	111,827,708	C/T	8.70x10^{-5}	-0.19	0.05	0.83	I	Y
2	rs2663966	145,563,703	C/T	8.98x10^{-5}	0.14	0.04	0.30	I	Y
1	rs12567277	4,558,398	C/T	9.09x10^{-5}	-0.13	0.03	0.16	G	Y
4	rs1462367	88,789,552	C/T	9.10x10^{-5}	0.15	0.04	0.34	G	Y
6	rs13208776	168,684,473	A/G	9.37x10^{-5}	-0.14	0.03	0.25	G	Y
18	rs937021	44,638,071	A/G	9.49x10^{-5}	0.13	0.03	0.60	G	Y
1	rs3007704	150,218,699	C/T	9.55x10^{-5}	-0.22	0.06	0.18	G	N
11	rs7951189	66,037,846	A/G	9.84x10^{-5}	-0.14	0.04	0.94	I	Y

As 88% of the most associated SNPs highlighted in Table 3.7 had been genotyped in the EngP2 and ScotP2 datasets, the decision to include these samples in the overall meta analysis was considered to increase power to detect new variants. The addition of EngP2 and ScotP2 would increase the sample size by 4,878 cases and 4,914 controls. Although these two datasets were genotyped for fewer SNPs (approximately 50,000 in total), the increase in power to detect common SNPs that were included outweighed the potential loss of a small number of truly associated variants of small effect that did not make it into the phase two genotyping SNP lists. The effect of adding in EngP2 and ScotP2 increases power to detect an associated variant with allele frequency of 1% and relative risk of 1.1 at a significance level of 0.05 in a joint analysis (taking into account that not all SNPs are genotyped in phase two) to 64% (or 82% in a one stage design). However, when the phase two datasets are left out the power drops to 47%.

3.7.1 The meta-analysis results including EngP2 and ScotP2

Using a fixed-effects model, the meta-analysis of EngP1, EngP2, ScotP1, ScotP2 and VQ58 (including the imputed SNPs) detected all of the regions previously identified and the nine SNPs that reached a formal level of significance (defined as 1x10^{-7}) all fell within this group. There is one associated SNP included here that has not been mentioned previously; rs3802842 (at 11q23.1, 110,676,919bp) was originally identified by Malcolm Dunlop's group in ScotP1/P2 (Tenesa *et al.* 2008) and the association was later replicated in EngP1/2 and the replication datasets (P=1.08x10^{-12}, OR=1.17, 95% CI 1.12-1.22)(Pittman *et al.* 2008).

The analysis identified 17 additional loci containing SNPs with P values of less than 1x10^{-4} (see Table 3.8). SNPs were followed up, by genotyping in the replication phase, if the effect was replicated in the COIN/NBS dataset. SNPs that failed this stage of replication are indicated in the table. In order to confirm the association with CRC risk, the five remaining loci were taken forward for genotyping in the replication phase.

Table 3.8 The meta-analysis results of four GWA studies including EngP2 and ScotP2

The table shows the top associated SNPs with P<1x10^{-4} from the meta-analysis, using a fixed effects model, of EngP1, ScotP1, VQ58 (plus imputed Hap550 SNPs), EngP2 and ScotP2. SNPs were analysed in the COIN/NBS dataset to determine whether they should be taken forward to full replication phase (the results are indicated in the replication column, FR=failed replication). For SNPs imputed in VQ58, Type=I. SNPs previously identified are shown in bold. The SNPs rs706771 and rs7898455 on chromosome 10 are in high r^2 (0.96 and 0.93, respectively) with the previously identified rs10795668. The 7 underlined SNPs were selected for further analysis. SE is the standard error for the beta (log OR) values for each SNP.

Chr.	SNP	Position (bp)	Allele A/B	P value	Beta (logOR)	SE	P het	Type	COIN/NBS replication
1	rs12037907	67,745,023	A/C	1.39x10^{-5}	0.15	0.04	0.12	I	FR
1	rs11805285	218,418,346	C/T	2.55x10^{-5}	-0.13	0.03	0.38	G	
1	rs11118515	218,540,647	A/G	2.19x10^{-5}	-0.14	0.03	0.09	G	
1	rs6691170	220,112,069	G/T	2.92x10^{-5}	0.09	0.02	0.54	G	
1	rs949618	220,180,563	A/C	1.95x10^{-5}	0.11	0.03	0.97	G	
1	rs6687758	220,231,571	A/G	2.41x10^{-6}	0.13	0.03	0.99	G	
1	rs12125368	220,265,295	A/C	8.71x10^{-6}	-0.12	0.03	1.00	I	
2	rs12471545	12,085,526	C/T	3.12x10^{-5}	-0.09	0.02	0.23	I	FR
2	rs11679483	161,232,395	A/G	2.46x10^{-5}	0.12	0.03	0.41	G	FR
3	rs10936599	170,974,795	C/T	1.97x10^{-6}	-0.12	0.03	0.28	G	
3	rs3772190	170,983,181	A/G	2.05x10^{-6}	0.12	0.03	0.32	I	
3	rs1997392	170,992,346	C/T	2.19x10^{-5}	-0.10	0.02	0.31	G	
3	rs6793295	171,001,149	C/T	2.88x10^{-5}	0.10	0.02	0.27	G	
3	rs11709840	171,052,935	A/C	2.44x10^{-5}	-0.10	0.02	0.32	I	
3	rs1920116	171,062,665	A/G	2.31x10^{-5}	0.10	0.02	0.40	G	
3	rs7647589	171,064,917	A/G	6.41x10^{-5}	-0.09	0.02	0.66	G	
4	rs7682616	342,955	A/G	2.52x10^{-5}	-0.09	0.02	0.96	G	FR
4	rs3946	357,927	A/G	2.80x10^{-5}	-0.09	0.02	0.98	I	
4	rs2604558	14,716,775	A/G	4.01x10^{-5}	-0.09	0.02	0.73	G	FR
8	rs1464327	61,159,909	C/T	8.11x10^{-5}	-0.09	0.02	0.33	G	FR
8	**rs16892766**	**117,699,864**	A/C	**4.52x10^{-11}**	**0.25**	**0.04**	**0.45**	**I**	
8	rs11986063	117,709,496	C/T	4.90x10^{-10}	0.23	0.04	0.46	G	

8	rs6983626	117,871,329	C/T	2.19x10^{-6}	0.17	0.04	0.78	G	
8	rs10505476	128,477,298	C/T	1.90x10^{-6}	0.12	0.02	0.44	G	
8	rs10808555	128,478,693	A/G	8.76x10^{-9}	0.13	0.02	0.84	G	
8	**rs6983267**	**128,482,487**	G/T	**3.30x10^{-14}**	**-0.16**	**0.02**	**0.02**	**G**	
8	rs10505473	128,487,118	C/T	1.77x10^{-5}	0.13	0.03	0.94	G	
8	rs7837328	128,492,309	A/G	1.49x10^{-12}	-0.15	0.02	0.55	I	
8	**rs7014346**	**128,493,974**	A/G	**1.81x10^{-14}**	**-0.17**	**0.02**	**0.50**	**G**	
9	rs182762	101,076,734	A/G	8.91x10^{-5}	0.14	0.04	0.19	G	FR
10	**rs706771**	**8,736,452**	A/G	**5.85x10^{-6}**	**0.10**	**0.02**	**0.34**	**G**	
10	**rs7898455**	**8,778,914**	G/T	**3.82x10^{-6}**	**-0.11**	**0.02**	**0.28**	**I**	
10	rs7069923	18,770,374	C/T	8.73x10^{-5}	0.09	0.02	0.28	G	FR
11	**rs3802842**	**110,676,919**	A/C	**1.44x10^{-10}**	**0.15**	**0.02**	**0.31**	**G**	
11	rs10749971	110,694,368	A/G	1.44x10^{-6}	0.11	0.02	0.86	G	
12	rs7138945	48,825,686	G/T	6.30x10^{-5}	-0.09	0.02	0.04	G	
12	rs1362983	48,900,974	A/G	2.63x10^{-6}	0.11	0.02	0.42	G	
12	rs11169335	48,922,631	A/G	3.76x10^{-6}	0.10	0.02	0.44	I	
12	rs6580742	49,014,078	C/T	5.93x10^{-6}	0.12	0.03	0.22	G	
12	rs7134595	49,016,725	C/T	2.08x10^{-6}	0.11	0.02	0.46	I	
12	rs7312252	49,030,438	C/T	2.22x10^{-6}	0.11	0.02	0.45	G	
12	rs12303082	49,040,830	G/T	2.13x10^{-6}	0.11	0.02	0.46	G	
12	rs11833608	49,043,895	A/G	2.27x10^{-6}	-0.11	0.02	0.48	G	
12	rs7136702	49,166,483	C/T	1.04x10^{-5}	0.10	0.02	0.92	G	
12	rs11169507	49,301,776	A/G	7.51x10^{-5}	0.10	0.02	0.48	I	
12	rs952318	49,315,533	A/C	7.05x10^{-5}	0.10	0.02	0.48	I	
12	rs12427378	49,360,466	C/T	6.42x10^{-6}	-0.10	0.02	0.67	G	
12	rs2139930	49,375,554	G/T	8.61x10^{-6}	0.10	0.02	0.69	G	
12	rs4307773	49,430,699	C/T	1.83x10^{-5}	0.09	0.02	0.79	G	
12	rs11169552	49,441,930	C/T	8.84x10^{-6}	-0.11	0.02	0.35	G	
12	rs1344958	49,483,984	C/T	1.18x10^{-5}	-0.11	0.02	0.45	G	
13	rs6491545	99,590,735	C/T	2.84x10^{-5}	-0.09	0.02	0.24	I	
13	rs684215	99,671,946	C/T	1.53x10^{-5}	-0.10	0.02	0.41	G	FR
13	rs7318781	99,676,821	C/T	6.86x10^{-6}	0.10	0.02	0.52	G	FR
14	**rs4444235**	**53,480,669**	C/T	**5.57x10^{-6}**	**-0.10**	**0.02**	**0.09**	**G**	
15	rs12438604	30,760,289	A/C	2.88x10^{-5}	-0.11	0.03	0.34	I	
15	**rs4779584**	**30,782,048**	C/T	**3.00x10^{-9}**	**0.16**	**0.03**	**0.13**	**G**	
15	rs11632715	30,791,539	A/G	2.59x10^{-7}	-0.11	0.02	0.89	G	
15	**rs10318**	**30,813,271**	C/T	**1.65x10^{-7}**	**0.14**	**0.03**	**0.25**	**G**	
15	rs1919360	30,830,747	C/T	8.81x10^{-6}	0.12	0.03	0.35	G	
15	rs782907	59,163,435	A/G	6.22x10^{-5}	-0.09	0.02	0.45	G	FR
16	rs1111720	67,256,387	C/T	5.61x10^{-6}	-0.12	0.03	0.82	I	
16	rs3114396	67,258,992	A/G	3.80x10^{-5}	0.10	0.02	0.33	G	
16	rs2902323	67,293,793	C/T	5.39x10^{-6}	-0.11	0.02	0.78	G	
16	**rs9929218**	**67,378,447**	A/G	**1.44x10^{-7}**	**0.13**	**0.02**	**0.77**	**G**	
16	**rs1862748**	**67,390,444**	C/T	**7.57x10^{-7}**	**-0.13**	**0.03**	**0.59**	**I**	

18	rs937021	44,638,071	A/G	6.93×10^{-7}	0.11	0.02	0.76	G	
18	**rs4939827**	**44,707,461**	C/T	**3.02×10^{-14}**	**0.16**	**0.02**	**0.11**	**G**	
18	rs4464148	44,713,030	C/T	5.87×10^{-7}	-0.12	0.02	0.07	G	
18	rs2337107	44,713,321	C/T	4.71×10^{-5}	0.09	0.02	0.28	G	
18	rs1316447	44,726,674	A/G	6.14×10^{-5}	-0.11	0.03	0.02	G	
19	**rs10411210**	**38,224,140**	C/T	**9.35×10^{-8}**	**-0.21**	**0.04**	**0.10**	**I**	
19	**rs7259371**	**38,226,481**	A/G	**1.06×10^{-5}**	**0.13**	**0.03**	**0.37**	**G**	
20	rs355527	6,336,068	C/T	2.11×10^{-7}	0.12	0.02	0.78	I	
20	**rs961253**	**6,352,281**	A/C	**2.60×10^{-7}**	**-0.13**	**0.03**	**0.90**	**I**	
20	rs4925386	60,354,439	C/T	1.22×10^{-5}	-0.10	0.02	0.39	G	
20	rs6061231	60,390,312	A/C	3.07×10^{-5}	0.10	0.02	0.31	G	
21	rs6517623	41,126,007	G/T	5.76×10^{-5}	-0.09	0.02	0.82	G	FR

We selected SNPs to follow up from these results on the basis of P value and independence from other nearby significant SNPs. In terms of the correlation between SNPs based on pair-wise LD by r^2, the SNPs on chromosome 12 are separated into two main blocks from which SNPs rs11169552 and rs7312252, which were chosen to take forward to replication (see Figure 3.15). Similarly, the SNPs on chromosome 1 (see Figure 3.16) are separated by r^2 into three groups from which rs11805285, rs6691170 and rs6687758 were chosen. The most strongly associated SNPs on chromosome 3 and 20 are both highly correlated with nearby identified SNPs and so only these two SNPs, rs10936599 and rs4925386 were taken forward (see Figure 3.17 and Figure 3.18, respectively).

Figure 3.15 The SNAP association plot and pair-wise r^2 for the chromosome 12 SNPs

The r^2 values indicated in the top plot are in relation to rs11169552. The pair-wise r^2 values shown in the lower plot were calculated using the EngP2 control genotypes. The labelled SNPs were chosen for further analysis in additional samples.

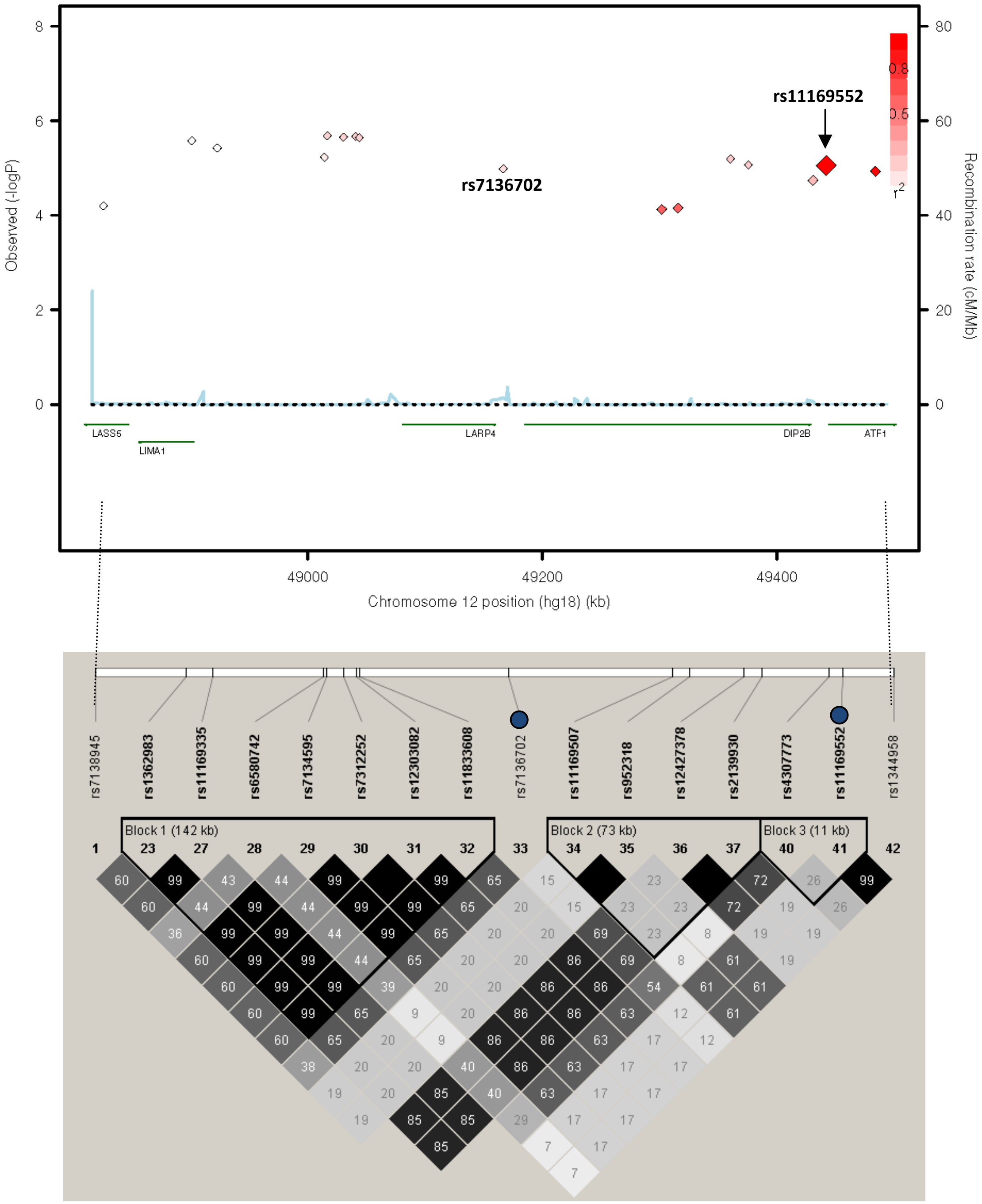


Figure 3.16 The LD for the chromosome 1 SNPs showing r^2 and D' values

The SNPs shown form three distinct groups by r^2 shown with black shading. D’ is shown in red.

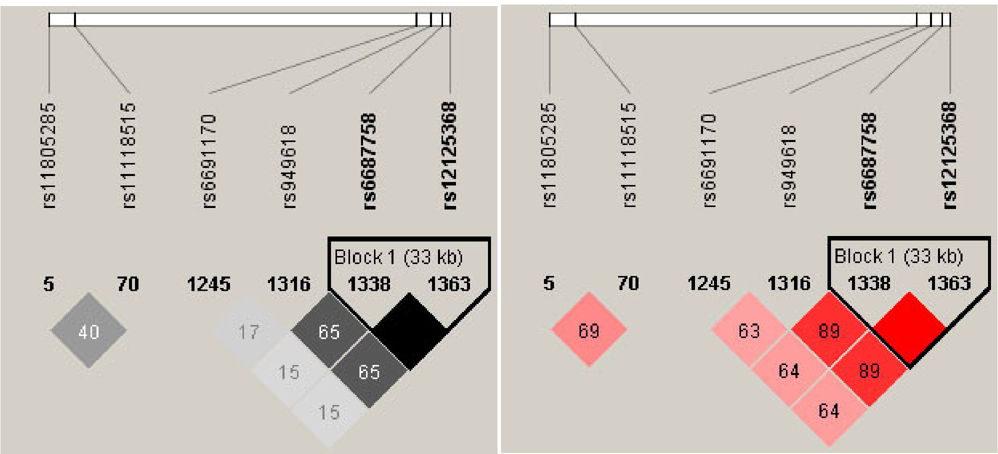


Figure 3.17 The LD for the chromosome 3 SNPs with $P<1x10^{-4}$ showing D’ and r^2

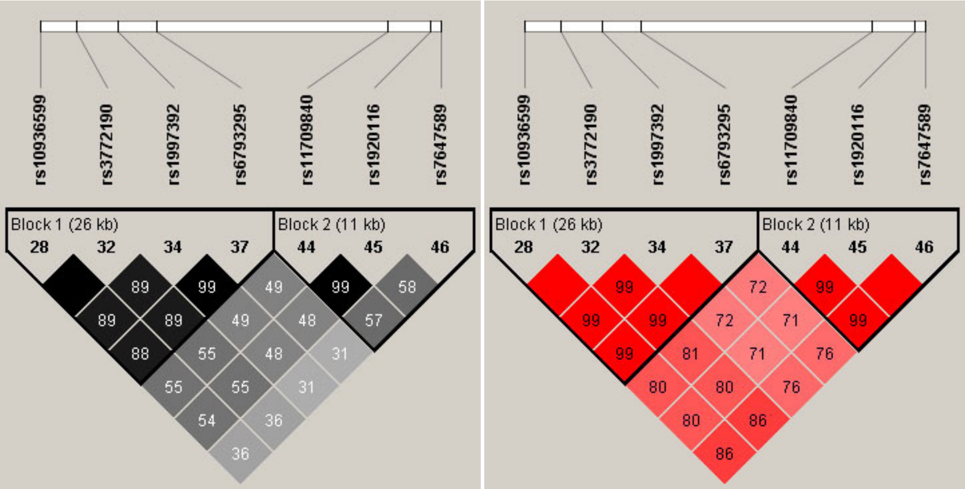
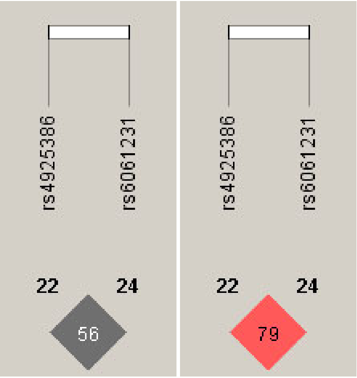


Figure 3.18 The LD for the chromosome 20 SNPs with $P<1x10^{-4}$ showing D’ and r^2



Once those SNPs in high pair-wise LD (r^2>0.7) were removed from each loci, seven SNPs with P values less than 5.0x10^{-5} were chosen for follow up replication studies in independent datasets. These SNPs (underlined in the table above) were rs6691170 (1q41), rs6687758 (1q41), rs11805285 (1q41), rs10936599 (3q26), rs7136702 (12q13), rs11169552 (12q13) and rs4925386 (20q13). All of these SNPs were actually genotyped in the VQ58 datasets rather than being imputed.

The replication phase (see Figure 3.14) consisted of samples from COIN/NBS, CORGI2BCD, Finland, EngP3, ScotP3, and SEARCH datasets, which are described fully in the materials and methods chapter. The combined replication phase meta analysis included a total of 38,232 samples. One of the seven SNPs failed to achieve significance in the replication phase, rs11805285 (original meta-analysis P=2.55x10^{-5}, while the combined replication P=3.90x10^{-5}), but associations were confirmed for the others (See Table 3.9). The data described in this section has now been incorporated into a recent publication (Houlston *et al.* 2010).

Table 3.9 The replication results summary for the seven detected SNPs

This table shows a summary of the results of the seven detected SNPs in each of dataset included in the analysis and contains the overall P value from the combined analysis of the replication datasets with the GWA study results, generated using a fixed effects model. The OR's were calculated with reference to the minor allele. The alleles are given as minor/major. The SNP rs11805285 was not genotyped in ScotP3 or SEARCH as the effect was not replicated in EngP3.

SNP	Position (bp)	Alleles	Dataset	P value	OR	95% CI	MAF aff	MAF ctrl
rs11805285 Chr. 1q41	218,418,346	T/C	EngP1	0.158	0.87	0.72-1.05	0.129	0.145
			EngP2	3.51x10^{-2}	0.89	0.80-1.05	0.125	0.139
			ScotP1	6.98x10^{-3}	0.78	0.65-0.93	0.126	0.155
			ScotP2	0.332	0.94	0.83-1.07	0.133	0.140

			COIN/NBS	2.33x10^{-2}	0.87	0.76-0.98	0.115	0.130
			VQ58	5.16x10^{-3}	0.82	0.72-0.94	0.119	0.141
			CORGI2bcd	0.481	0.92	0.74-1.15	0.133	0.143
			FCCPS	0.382	0.90	0.71-1.14	0.081	0.090
			EngP3	0.158	1.08	0.97-1.22	0.122	0.114
			Combined Analysis	**3.90x10^{-5}**	**0.92**	**0.88-0.96**		
rs6691170	220,112,069	T/G	EngP1	2.06x10^{-2}	1.17	1.02-1.34	0.378	0.341
Chr. 1q41			EngP2	3.02x10^{-3}	1.12	1.04-1.21	0.384	0.357
			ScotP1	7.42x10^{-2}	1.13	0.99-1.28	0.374	0.347
			ScotP2	0.514	1.03	0.94-1.13	0.358	0.351
			COIN/NBS	4.44x10^{-2}	1.09	1.00-1.19	0.384	0.364
			VQ58	9.10x10^{-2}	1.08	0.99-1.19	0.382	0.363
			CORGI2bcd	0.677	1.03	0.88-1.21	0.357	0.349
			FCCPS	5.20x10^{-2}	1.15	1.00-1.32	0.388	0.356
			EngP3	3.90x10^{-3}	1.12	1.04-1.21	0.378	0.352
			ScotP3	0.722	0.98	0.85-1.12	0.361	0.367
			SEARCH	3.46x10^{-3}	1.14	1.04-1.24	0.391	0.360
			Combined Analysis	**4.12x10^{-10}**	**1.06**	**1.04-1.08**		
rs6687758	220,231,571	G/A	EngP1	0.254	1.10	0.94-1.29	0.210	0.195
Chr. 1q41			EngP2	5.58x10^{-3}	1.14	1.04-1.25	0.215	0.194
			ScotP1	4.60x10^{-2}	1.17	1.00-1.36	0.222	0.196
			ScotP2	2.41x10^{-2}	1.13	1.02-1.26	0.211	0.191
			COIN/NBS	6.66x10^{-3}	1.15	1.04-1.28	0.212	0.190
			VQ58	3.82x10^{-2}	1.13	1.01-1.26	0.216	0.197
			CORGI2bcd	0.302	1.10	0.91-1.33	0.211	0.195
			FCCPS	0.161	1.11	0.96-1.30	0.280	0.258
			EngP3	0.268	1.05	0.96-1.15	0.199	0.191
			ScotP3	0.594	0.96	0.82-1.12	0.216	0.224
			SEARCH	4.38x10^{-3}	1.16	1.05-1.29	0.211	0.187
			Combined Analysis	**1.18x10^{-9}**	**1.09**	**1.06-1.12**		
rs10936599	170,974,795	T/C	EngP1	2.23x10^{-3}	0.78	0.67-0.92	0.202	0.244
Chr. 3q26			EngP2	2.10x10^{-2}	0.90	0.83-0.98	0.229	0.248
			ScotP1	5.92x10^{-3}	0.82	0.70-0.90	0.226	0.264
			ScotP2	0.158	0.93	0.84-1.03	0.240	0.254
			COIN/NBS	0.188	0.94	0.85-1.03	0.236	0.248
			VQ58	7.13x10^{-2}	0.91	0.81-1.01	0.227	0.245
			CORGI2bcd	0.117	0.87	0.73-1.04	0.227	0.253
			FCCPS	6.99x10^{-2}	0.87	0.75-1.01	0.256	0.284
			EngP3	0.750	0.99	0.91-1.07	0.236	0.238
			ScotP3	0.181	0.90	0.77-1.05	0.222	0.241
			SEARCH	7.26x10^{-2}	0.91	0.83-1.01	0.233	0.249
			Combined Analysis	**2.51x10^{-8}**	**0.93**	**0.91-0.96**		
rs7136702	49,166,483	T/C	EngP1	0.126	1.11	0.97-1.27	0.377	0.353
Chr. 12q13			EngP2	4.28x10^{-2}	1.08	1.00-1.17	0.367	0.348
			ScotP1	6.22x10^{-2}	1.13	0.99-1.29	0.376	0.348
			ScotP2	6.68x10^{-2}	1.09	0.99-1.19	0.381	0.361
			COIN/NBS	0.326	1.04	0.96-1.14	0.362	0.352
			VQ58	1.03x10^{-2}	1.13	1.03-1.24	0.378	0.350
			CORGI2bcd	0.165	1.12	0.96-1.31	0.388	0.362

			FCCPS	6.25×10^{-2}	1.15	0.99-1.33	0.321	0.291
			EngP3	0.138	1.06	0.98-1.14	0.368	0.355
			ScotP3	0.163	1.11	0.96-1.28	0.391	0.367
			SEARCH	8.96×10^{-2}	1.08	0.99-1.18	0.370	0.352
			Combined Analysis	**2.01×10^{-8}**	**1.06**	**1.04-1.08**		
rs11169552	49,441,930	T/C	EngP1	0.129	0.89	0.77-1.03	0.239	0.260
Chr. 12q13			EngP2	0.201	0.95	0.87-1.03	0.260	0.270
			ScotP1	5.09×10^{-2}	0.87	0.75-1.00	0.251	0.279
			ScotP2	8.77×10^{-2}	0.92	0.83-1.01	0.257	0.273
			COIN/NBS	0.511	0.97	0.88-1.06	0.264	0.270
			VQ58	3.35×10^{-4}	0.82	0.74-0.92	0.233	0.269
			CORGI2bcd	0.251	0.90	0.76-1.08	0.250	0.270
			FCCPS	1.55×10^{-5}	0.74	0.64-0.85	0.336	0.408
			EngP3	3.72×10^{-3}	0.89	0.82-0.96	0.255	0.278
			ScotP3	0.212	0.88	0.71-1.08	0.244	0.270
			SEARCH	0.135	0.93	0.85-1.02	0.255	0.269
			Combined Analysis	**9.40×10^{-11}**	**0.92**	**0.90-0.95**		
rs4925386	60,354,439	T/C	EngP1	3.37×10^{-3}	0.81	0.70-0.93	0.273	0.316
Chr. 20q13			EngP2	0.113	0.94	0.87-1.02	0.297	0.311
			ScotP1	1.77×10^{-2}	0.85	0.74-0.97	0.283	0.318
			ScotP2	2.63×10^{-2}	0.90	0.82-0.99	0.292	0.315
			COIN/NBS	0.977	1.00	0.92-1.09	0.308	0.307
			VQ58	9.76×10^{-2}	0.92	0.83-1.02	0.306	0.324
			CORGI2bcd	0.197	0.90	0.76-1.06	0.292	0.315
			FCCPS	0.729	0.97	0.84-1.13	0.297	0.302
			EngP3	1.41×10^{-5}	0.84	0.78-0.91	0.289	0.326
			ScotP3	6.40×10^{-2}	0.87	0.76-1.01	0.295	0.324
			SEARCH	8.56×10^{-3}	0.89	0.81-0.97	0.289	0.314
			Combined Analysis	**6.81×10^{-11}**	**0.93**	**0.91-0.95**		

3.8 Network analysis of the candidate genes associated with the identified SNPs

I have used the web based program GLIDERS (Genome-wide Linkage Disequilibrium Repository and search engine available at http://mather.well.ox.ac.uk/GLIDERS/) to determine which SNPs in HapMap 2/3 are in LD (D’>0.7) with each of the most strongly associated SNPs and used this as a proxy to indicate the boundaries of the LD ‘block’. An analysis can then be performed on the known genes that reside with this block, around each of the significant SNPs, to determine if there are any overlaps in the pathways or overall networks that these genes work in and to identify possible candidate genes. The genes that are located within the block indicated by the SNPs, which were in LD with the test SNP (see Table 3.10), have been identified using the UCSC genome browser (http://genome.ucsc.edu/). The gene IDs were found using the online NCBI database Entrez Gene (http://www.ncbi.nlm.nih.gov/gene).

Table 3.10 Identified associated SNPs with LD region and included genes

The table of all identified associated SNPs showing the flanking region defined by LD (D’) with other SNPs in HapMap and the genes that reside within this region. The SNP rs1295371 was not contained in the GLIDERS database and so a region is not given.

SNP	Chr	Position	Region of LD		Genes
		(bp)	Start (bp)	End (bp)	
rs6983267	8q24.21	128,482,487	128,476,625	128,522,947	POU5F1B, DQ515898, DQ515897, DQ515899
rs10505477	8q24.21	128,476,625	128,477,298	128,497,243	As above
rs4939827	18q21.1	44,707,461	44,705,144	44,708,046	SMAD7
rs1295371	18q21.1	44,707,927	-	-	SMAD7
rs4464148	18q21.1	44,713,030	44,707,927	44,726,674	SMAD7
rs4779584	15q13.3	30,782,048	30,782,590	30,840,760	GREM1, DRM
rs16892766	8q23.3	117,699,995	117,678,424	117,884,259	EIF3H, UTP23, RAD21
rs10795668	10p14	8,741,225	8,683,135	8,778,914	BX100511
rs3802842	11q23.1	110,676,919	110,630,722	110,685,600	C11orf53, C11orf92, LOC120376, POU2AF1
rs9929218	16q22.1	67,378,447	67,010,401	67,396,803	CDH1, CDH3, ZFP90 KIAA1954, SMPD3
rs1862748	16q22.1	67,390,444	67,127,146	67,396,803	As above
rs4444235	14q22.2	53,480,669	53,457,559	53,501,325	BMP4
rs10411210	19q13.1	38,224,140	38,225,132	38,308,096	RHPN2, GPATCH1
rs7259371	19q13.1	38,226,481	38,208,281	38,314,782	RHPN2, GPATCH1 DKFZp761L1918
rs961253	20q12.3	6,352,281	6,231,026	6,354,102	BMP2, FERMT1
rs355527	20q12.3	6,336,068	6,231,026	6,358,854	AI971377 (EST)
rs6691170	1q41	220,112,069	220,033,404	220,206,329	AW590668 (EST) AA782022 (EST) AI732358 (EST)
rs6687758	1q41	220,231,571	220,122,380	220,295,514	As Above

rs10936559	3q26	170,974,795	170,899,263	171,064,917	TERC, MYNN, ARPM1, LRRC34, LRRIQ4, LRRC31
rs7136702	12q13	49,166,483	48,793,976	49,507,394	C12ORF62, LASS5, LIMA1, LARP4, DIP2B, ATF1
rs11169552	12q13	49,441,930	49,110,145	49,504,590	LARP4, DIP2B, ATF1
rs4925386	20q13	60,354,439	60,320,976	60,404,070	LAMA5, RPS21, CABLES2

This gene list was used as a query in GeneGo (http://www.genego.com) to determine shared networks and pathways, determined based on enrichment for the genes in the list (the network and pathway maps are given in the appendix). Several pathways were identified and the top three pathways ordered by gene enrichment were BMP signalling, Cadherin cell adhesion and WNT signalling, but many of the pathways in the list are relevant for cancer predisposition (Table 3.11). However, these pathways cover 10 of the above predisposition loci and only the top four in the list include genes from more than one of the loci. It should be noted that the genes *CDH1*, *ATF1* and *SMAD7* appear in multiple pathways and are involved in many interactions and so it is difficult to determine which pathways are involved, as all are likely to have a role in cancer risk.

Table 3.11 Molecular pathways or processes for the identified susceptibility loci

The gene list in Table 3.10 was used to determine the molecular pathways in which the protein products of these candidate genes function. The pathways are listed in order of enrichment for the gene list, although *CDH1*, *ATF1* and *SMAD7* appear in multiple pathways.

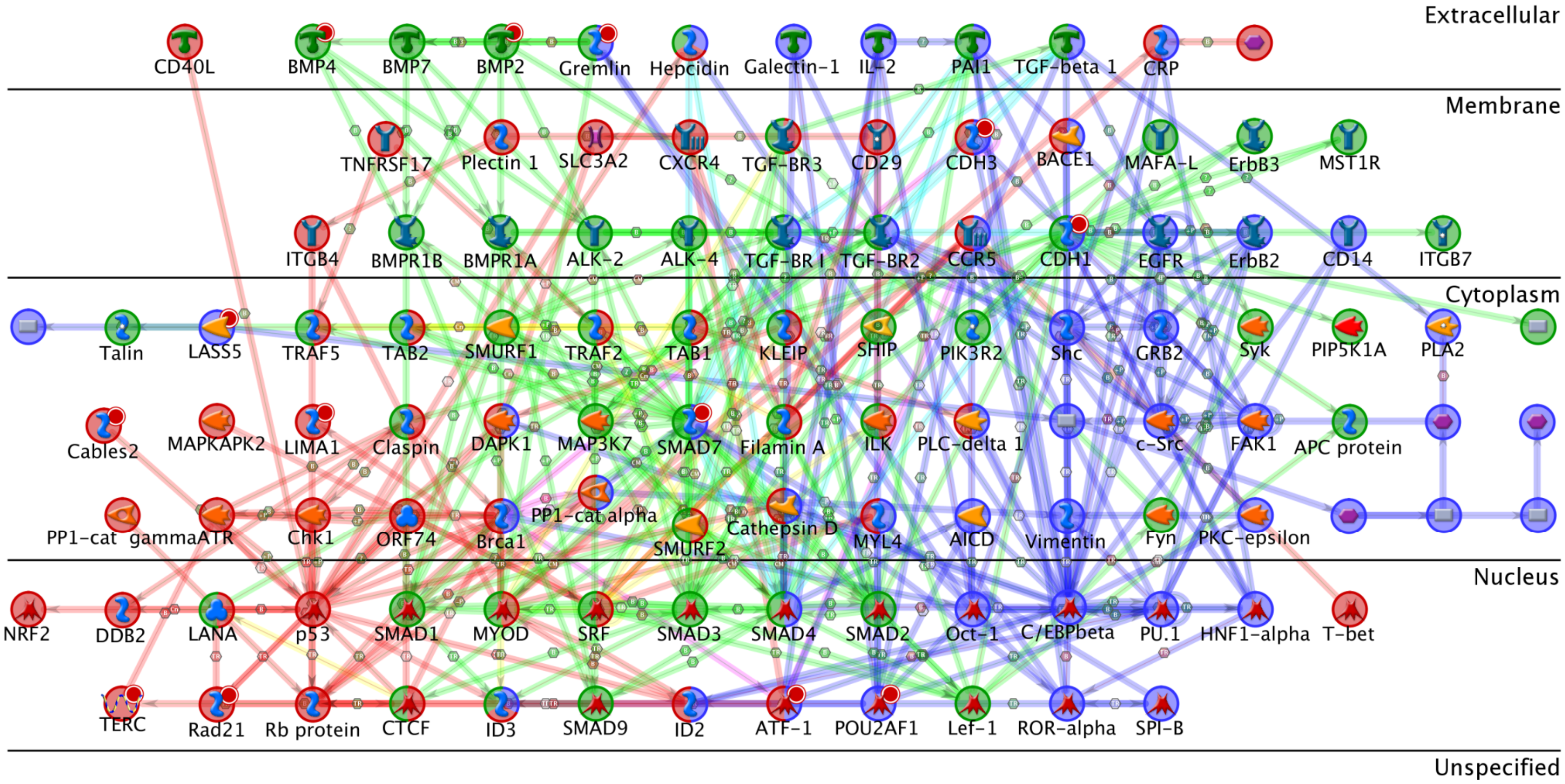
Pathway / Process	General function	Candidate Genes
BMP signalling	Development	BMP2, BMP4, SMAD7, GREMLIN
Cadherin mediated cell adhesion	Cell Adhesion	CDH1, CDH3
WNT signalling pathway	Development	CDH1, BMP4
Sphingolipoid Metabolism	Metabolism	LASS5, SMPD3

CDK5 in cell adhesion	Cell adhesion	CDH1
Hypoxia induced EMT in cancer and fibrosis	Hypoxia	CDH1
MicroRNA dependent inhibition of EMT	Development	CDH1
Beta-2 adrenergic dependent CFTR expression		ATF1
RhoB regulation pathway	G-Protein Signalling	RHPN2
NOTCH induced EMT	Development	CDH1
Sister chromatid cohesion	Cell cycle	RAD21
TGFB1 dependent inhibition of CFTR expression		SMAD7
Regulation of initiation of translation	Translation	EIF3H, RPS21
BRCA1 and BRCA2 in DNA repair	DNA damage	ATF1
BRCA1 – transcription regulator	DNA damage	ATF1
HGF dependent inhibition of TGFB induced EMT	Development	SMAD7
PACAP signalling in neural cells	Development	ATF1
Histamine H1 receptor signalling in the interruption of cell barrier integrity.	Cell adhesion	CDH1
Thrombopoietin regulated cell processes	Development	ATF1

Many of the genes in the gene list can be networked showing that although they are not necessarily in common pathways there are downstream interactions between members of these pathways that link many of the genes close to associated SNPs (Figure 3.19). These analyses highlight many additional genes that may influence CRC risk through disruption of the same network or process. Of course, this is a crude analysis and more work would be required. I have only included the genes that are on the same chromosome as the identified SNP and also the gene list includes all genes in the region defined by SNPs in LD with the associated SNP; I did not prune my gene list by function.

Figure 3.19 Network of genes within the region of identified SNPs

This plot illustrates the overall network produced when the genes for the top three subnetworks, ordered by relevance to the candidate genes listed in Table 3.10, are combined. The network includes 12 of the candidate genes and the intermediate nodes that link them. The diagram was produced in the pathway analysis software Genego (MetaCore), using the build network option, using a minimum of 50 nodes. Nodes labelled with a red dot are present in the candidate gene list.



3.9 Discussion

The results discussed in this chapter summarise the findings of the CRC GWA study and show that we have robustly identified 14 loci significantly associated with CRC risk. Together these loci are estimated to increase the risk of CRC by approximately 5 fold. A recent investigation into the combined effect of the first ten associated SNPs described in this Chapter has shown that each additional risk allele carried by an individual increased the likelihood of having an affected first degree relative by 1.16 (95% CI of 1.04-1.30)(Niittymaki *et al.* 2010).

Perhaps surprisingly, the SNPs identified by this GWA study do not tag mismatch repair genes or other known candidate genes for increased CRC risk. However, as determined in Section 3.7, most are within LD blocks that contain genes that could have a functional effect on cancer susceptibility or function within pathways or networks containing for example, development pathways, such as BMP and WNT signalling, and pathways, such as Cadherin mediated cell adhesion. All of these pathways include good candidate genes for CRC risk and warrant further investigation.

These findings demonstrate that the common-disease common-variant hypothesis is correct, but also that there remain variants that lie undiscovered by this approach. It is probable that the true answer lies in a combination of common and rare variants and additional chromosome aberrations. The common-disease rare-variant hypothesis states that a significant proportion of susceptibility may be due to low frequency variants with a moderate effect. The GWA study is not designed to detect rare variants and the SNPs included on the array used in this analysis is optimised to tag common variants with MAF greater than 10%. Therefore, alternative methods are required, such as sequencing, to determine the effect of these rare variants. Next generation sequencing projects are already underway for the associated loci to determine locations of mutations in patients and identify rare variants for analysis.

The meta-analysis studies have illustrated that SNPs low down in the ranked list of associations from any individual GWA study can still represent real associations that are later replicated when the power of the study is increased by, for example, the addition of more samples. This can mean that the usual method of prioritising SNPs for follow up, which involves taking forward only the most significant SNPs in the discovery phase and attempting to replicate the results in additional studies, can result in genuinely associated SNPs further down the list being overlooked. It is better to combine GWA studies by meta-analysis to ensure that all SNPs on the genotyping arrays, including those with small effects, are included in the analysis to increase the chances of detection. Most of the effect sizes of SNPs detected in this study were below 1.2, and the meta-analysis of the main GWA studies elucidated many SNPs with much smaller effect sizes that would not have been detected from the initial GWA alone.

The quality control methods employed prior to combining the results from multiple GWA studies have highlighted the issues faced when incorporating additional studies into meta-analyses and importance of PCA analysis and other methods to identify population structure and the presence of duplications and related samples across studies collecting for the same disease from the same population. Some of these issues could be missed when only utilising additional datasets for replication analyses of single SNPs, as whole genome data is generally not exchanged and so this type of analysis is difficult.

Chapter 4. Imputation of SNP genotypes and additional analyses on disease-associated loci

4.1 Introduction

The previous Chapter provided a summary of the work undertaken as a large collaborative project where everyone involved contributed to the research. In this Chapter, I discuss the aspects of the research that were my personal focus, the majority of which is based around the imputation of un-typed or 'missing' SNPs and how the data were used to enhance the GWA study.

4.2 An introduction to Imputation

The analyses performed as part of the GWA study involved testing association with disease at individual markers that act as tags or predictors of untyped common variation. The advantages of being able to indirectly detect disease-associated SNPs over the whole genome by genotyping a subset of tag SNPs is a limitation in terms of fine-mapping identified disease loci. The SNPs genotyped may not be the SNPs in highest LD with the causal variant or the most strongly associated variants for the disease being studied and, since they are a small fraction of the known variants, are unlikely to be causal variants.

Therefore, to better refine the analysis and facilitate fine-mapping, one can increase the number of SNPs analysed by incorporating additional known variants (described in public databases) by using the genotyped SNPs to predict (or 'impute') the genotypes of those variants not included on the SNP arrays. In the same way that haplotype analysis, as a multipoint method, should improve the detection of unobserved variants compared to single marker tests, imputation is a method for predicting the unobserved variation so that it may be ‘directly’ analysed for association. This should aid fine-mapping methods that solely rely on LD to locate the causal variant as it applies a model to the data to accurately estimate the level of uncertainty (Zollner and Pritchard 2005).

Owing to the small effect sizes of SNPs identified thus far, large numbers of samples are required to provide sufficient power to detect additional susceptibility variants. These are usually achieved through collaborations between groups with similar collections of samples. However, there are a various genotyping arrays available, which include different panels of SNPs and provide varying coverage of the genome. Therefore, studies either need to be genotyped using the same arrays or have sufficient overlapping SNPs to enable meta-analysis. Imputation provides a solution to these limitations by predicting the genotypes at untyped SNPs, which can then be used to provide the overlap required for meta-analyses with other datasets. Imputation has been used in several published GWA studies to refine genotyped SNP results and fine-map associated regions, including the WTCCC GWA study (Wellcome Trust Case Control Consortium 2007) and a separate study to refine the association of the 15q25 locus with smoking quantity using the 1KGP data (Liu *et al.* 2010).

4.2.1 The IMPUTE imputation program

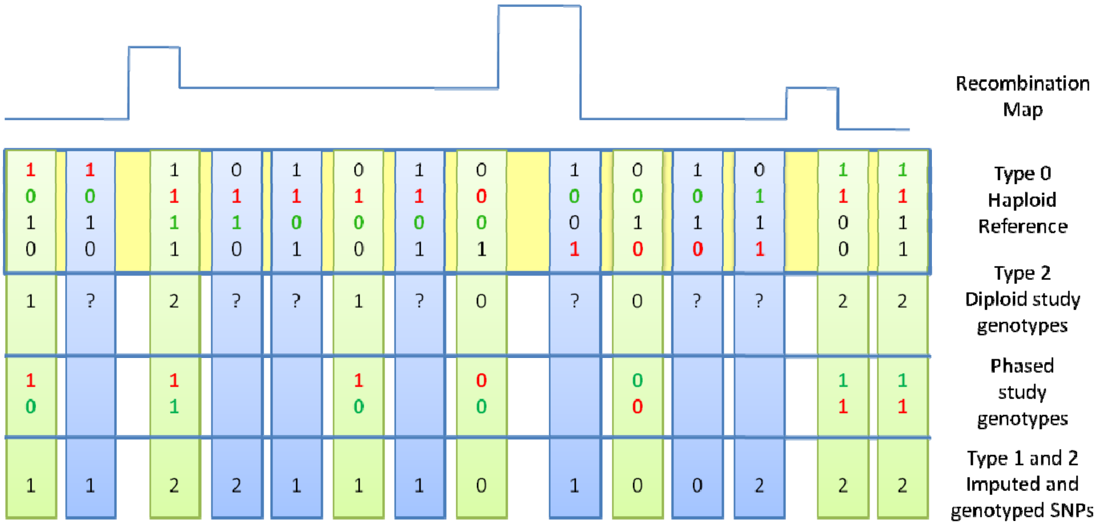
There are a number of programs available for the imputation of missing genotypes, but I chose to use IMPUTE (Marchini *et al.* 2007). This decision was made on the basis of improved performance over other available programs and local expertise (Pei *et al.* 2008; Howie *et al.* 2009). IMPUTEv1 predicts missing genotypes using a combination of the genotypes of the SNPs on the array, a reference panel of phased haplotypes such as HapMap (which includes the genotyped SNPs and those to be imputed), a fine-scale genome wide recombination map and a population genetics model to infer allelic correlation from the reference panel (Marchini *et al.* 2007). This facilitates the imputation of all SNPs included in the reference panel provided there is sufficient LD between the genotyped SNPs in the study and the untyped SNPs to be imputed.

Impute uses a population genetics model to assign more weight to genotypes that follow local LD patterns and utilise genotypes from all genotyped SNPs that are in LD with the untyped SNP. The model is not valid over large distances and so imputing a large region of the genome at one time can lead to poor quality results. Therefore, chromosomes need to be split into smaller chunks for imputation (Howie *et al.* 2009).

IMPUTE is based on an extension of the Hidden Markov Model, developed for uses such as modelling LD and estimating recombination rates. Each individual is phased and modelled against the reference panel to form a mosaic of haplotypes, owing to the unobserved SNPs in the study data. The haplotypes in the reference panel are then used to impute the untyped SNP genotypes (see Figure 4.1). A key assumption is that the haplotypes that match at typed SNPs will also match at the untyped SNPs. The phasing accuracy of the genotyped inference panel (the study genotypes) determines how well the haplotypes match the reference panel haplotypes and hence how well the imputation will perform.

Figure 4.1 An overview of imputation

This schematic shows a simplified overview of what impute does based on information provided on the program website (https://mathgen.stats.ox.ac.uk/impute/impute.html) and shows the distinction between the reference panels and the inference panels. Type 1 SNPs are present in the reference panel only and type 2 SNPs are genotyped in the inference (study) sample and the reference panel. If an additional reference panel is included, the SNPs in this panel are labelled type 0. The bottom section of this diagram shows the results of the imputation (in reality the genotype in the output is a probability distribution e.g. 0, 0, 1). The red and green figures indicate the line the program has taken through the haplotypes of the reference panel to enable imputation, which allow switches between haplotypes dependent on the recombination map.


Recombination Map
Type 0 Haploid Reference
Type 2 Diploid study genotypes
Phased study genotypes
Type 1 and 2 Imputed and genotyped SNPs

Type 0 Haploid Reference	1 0 1 0	1 0 1 0	1 1 1 1	0 1 1 0	1 1 0 1	0 1 0 0	1 1 0 1	0 0 0 1	1 0 0 1	0 0 1 0	1 0 1 0	0 1 1 1	1 1 0 0	1 1 1 1
Type 2 Diploid study genotypes	1	?	2	?	?	1	?	0	?	0	?	?	2	2
Phased study genotypes	1 0		1 1			1 0		0 0		0 0			1 1	1 1
Type 1 and 2 Imputed and genotyped SNPs	1	1	2	2	1	1	1	0	1	0	0	2	2	2

IMPUTEv1 predicts untyped genotypes one sample at a time based on the reference panel and uses this to integrate over phase uncertainty. The genotypes of one individual do not influence the imputation of another. However, IMPUTEv2 incorporates information from all available genotypes, not just those of the individual being imputed, and uses this to predict the phase of the study genotypes with more accuracy. A Monte Carlo Markov Chain (MCMC) algorithm is used to integrate over phase uncertainty by performing multiple iterations of two steps and then averaging over the resulting genotype probabilities. In step one, the observed genotypes are phased, any missing genotypes are imputed, and the phase data over all study samples are combined. In step two, the genotype probabilities for the untyped SNPs for each of the inferred haplotypes are imputed in each individual separately (Howie *et al.* 2009). This version also allows for the use of more than one reference panel and accepts unphased genotypes as reference panels, such as GWA study data, which are phased as part of the imputation process.

The study samples described in this thesis are of North European descent and so the main reference panel used for imputation was the CEU HapMap phase II dataset, which includes 30 parent-offspring trios from Utah, USA with Northern or Western European descent that were part of the CEPH collection. The haplotypes of these individuals are the closest population match, within HapMap, to the study samples.

4.2.2 Aspects that affect imputation accuracy

The accuracy of the imputation is greatly influenced by the quality of the study genotype data, but also by the density of the genotyped SNPs and whether they are tagging SNPs. All of the study datasets used in this GWA study have been genotyped on Illumina arrays (HumanHap300, 550 and 1M), which have the advantage that the SNPs are tagging SNPs chosen to best tag the known common variation in the genome. Based on the HapMap phase II data, the Illumina HumanHap550 array provides a genomic coverage of 87% at $r^2 \geq 0.8$, while the genomic coverage is 65% for the Affymetrix 5.0 and 80% for the Affymetrix 6.0 arrays (Anderson *et al.* 2008).

Genome wide arrays based on tagging SNPs are better suited to imputation as it is more likely that there will be genotyped SNPs in LD with those to be imputed. This is in contrast to the early Affymetrix arrays where the SNPs were chosen based on equal spacing across the genome. However, the Hap550 and Hap300 SNP arrays were not designed to tag low MAF SNPs (less than 0.05), which can lead to poor imputation of these SNPs. SNPs that have a low minor allele frequency and poor representation of this allele in the reference population will generally not impute with high accuracy, but instead all genotypes will often be homozygous for the common allele. This situation can be improved by increasing the number of samples in the reference panel, which increases the chance that the minor allele of the SNP will be well represented.

Other aspects affecting imputation accuracy include LD structure, the size of the reference panel, the density of markers and the population of the reference panel. The population genetic model should work best if the study data is from the same population as the reference panel owing to shared LD patterns. However, it has been shown recently that if individuals from all populations within HapMap are included in the reference panel, imputation quality for SNPs with a MAF less than 5% can be improved (Marchini and Howie 2010).

4.2.3 Quality control

IMPUTE produces an information score to give a relative measure of how well each SNP imputed. The information score is calculated using an established missing data likelihood theory to calculate the observed data information about a parameter. If the information score equals 1, the genotype prediction is 100% certain, an information score of close to 0 indicates that there is no confidence in the prediction. The information score can be used as a quality control to reject SNPs that do not impute well enough. For instance, an information score of 0.5 for a SNP indicates that the information provided equates to perfect genotype data in a sample half the size as that used in the original study (i.e. the information score x 100%). Therefore, the chosen threshold should take the sample size of the study into account. The information score threshold used was 0.5, as used in other GWA studies (Liu *et al.* 2010). The genotype probabilities produced by IMPUTE also provide a measure of the confidence in the predicted genotype, and so for this study I only included SNPs where at least 95% of samples achieved a maximum probability of 0.9.

Further investigation into how well the dataset is imputing can be performed using the screen messages captured during the program run and the concordance scores to assess how well the genotyped SNPs imputed when removed one at a time from the study panel.

4.2.4 SNPTEST for association analysis of imputed SNPs

The main output from IMPUTE contains the genotypes, for both typed and imputed SNPs, given as three probabilities for each individual, one for each possible genotype (AA, AB and BB). There are several ways in which the imputed genotype probabilities can be used to perform an association analysis; however the uncertainty in the genotype (as defined by the probability distribution) and the missing data need to be taken into account. One could set a threshold at which to call a genotype using the maximum genotype probability and only call SNPs that pass this threshold, over 0.9 for example, (best guess genotypes). This method does not take into account the uncertainty of the genotype and will only work well when the genotype certainty is very high. If the genotype certainty is low, there will be a high degree of missing data or uncalled genotypes that could skew the results. Furthermore, as heterozygotes are called with less certainty than non-heterozygotes, they will be uncalled more often using best guess genotypes which could introduce bias against heterozygotes.

An alternative method is to make an estimate of the expected genotype counts by summing the probabilities for each genotype, thus incorporating all the information about each SNP, but this still does not fully take uncertainty into account.

However, a better method, and the one I used to deal with this data, was implemented in SNPTESTv1 (the 'proper' command), which uses the output from IMPUTE to perform association analyses on the genotypes while taking the uncertainty of the imputed genotype probabilities into account. SNPTESTv1 uses the statistical theory for dealing with missing data using the observed data likelihood, where "the contribution of each possible genotype is weighted by its imputation probability" (Marchini and Howie 2010). This likelihood is used in a score test, which attempts to maximise the likelihood, to test for association. However, in situations where there is a high uncertainty in the genotype, small sample size and low allele frequency, this method of dealing with uncertainty can result in artificially low P values, although this is not a problem if low MAF SNPs are removed and has been improved in SNPTESTv2.

4.3 The imputation of Hap550 SNPs from the Hap300K genotypes

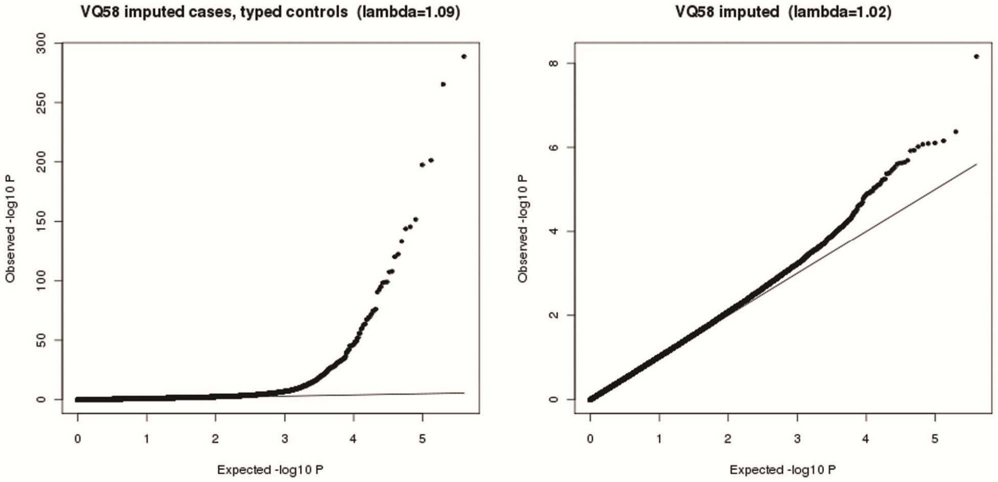
The cases for the VQ58 dataset were genotyped using the Illumina HumanHap300, whereas EngP1 and ScotP1 were genotyped using the Illumina HumanHap550 arrays. Although all the associated SNPs discovered so far were genotyped on both the Hap300 and the Hap550, to ensure that no signals were missed, I endeavoured to predict the genotypes of the approximately 200,000 untyped SNPs in VQ58 using the program IMPUTE and the HapMap phase II CEU samples as a reference panel. This allowed better overlap of SNPs with the datasets genotyped on the Hap500 SNP chips and presented the opportunity to explore the utility of this technique in our own datasets. For this analysis, I used IMPUTEv1 for imputation to HapMap phase II and IMPUTEv2 to incorporate additional reference panels, such as the unphased genotypes from the EngP1 controls and the 1000 genomes project pilot data.

The VQ58 dataset was prepared by performing basic quality control procedures of removing SNPs that were not called in at least 95% of samples and removing samples that did not achieve a 95% call rate. After removal of SNPs failing HWE (at P less than $1x10^{-6}$), the association results for the genotyped SNPs, based on allelic P value, showed no evidence of inflation and lambda was calculated at 1.02 (see Figure 3.1 in chapter 3). The association results revealed 35 SNPs with P values less than $1x10^{-4}$, but no SNPs achieved genome-wide significance based on a Bonferroni correction (defined as $P \leq 1x10^{-7}$).

As the controls for the VQ58 dataset were genotyped on the Illumina HumanHap1M SNP array, we initially did not want to ignore the information from these genotyped SNPs and decided to impute only the cases. However, from the results of the association analysis, it quickly became apparent that using directly typed genotypes in controls and imputed genotypes in cases leads to several issues. The most prominent is shown by the inflation of association test statistics in certain SNPs caused by a marked difference in allele frequency between the imputed cases and the genotyped controls (see Figure 4.2). Perhaps expectedly, the allele frequencies in these imputed SNPs were comparable to those given in dbSNP, which is calculated from the HapMap CEU samples. Imputation can highlight errors in genotyping and, therefore, differences in allele frequency could be caused by a genotype calling error in the genotyped samples. However, this could also indicate a genotyping problem with the HapMap or a genuine difference, affecting certain SNPs, between our GWA study datasets and the small number of CEU samples in the HapMap.

Figure 4.2 The QQ plot comparison for VQ58 including imputed data

A comparison of the inflation in P value observed, using the same group of SNPs, when the cases are imputed and the controls are typed (left) and when both VQ58 cases and controls are imputed (right).



To attempt to account for this, I used stringent quality control for the imputed genotypes. I initially used a 99% call rate (at a maximum probability threshold of 0.9) and an impute info score of greater than 0.4, but also removed SNPs that differed in allele frequency by more than 0.1 between genotyped and imputed datasets. Additionally, SNPs with a minor allele frequency of less than 0.05 in either the genotyped controls or the imputed cases were removed from the analysis. However, the overly stringent quality control measures to overcome these issues left just 41,098 SNPs for analysis. Even with these criteria it was clear, from the large number of significant P values for those SNPs imputed in the cases and genotyped in the controls, that treating the cases and controls differently introduces a bias in the results causing enough of a difference to inflate the test statistics. In this situation it is very difficult to determine true associations with disease.

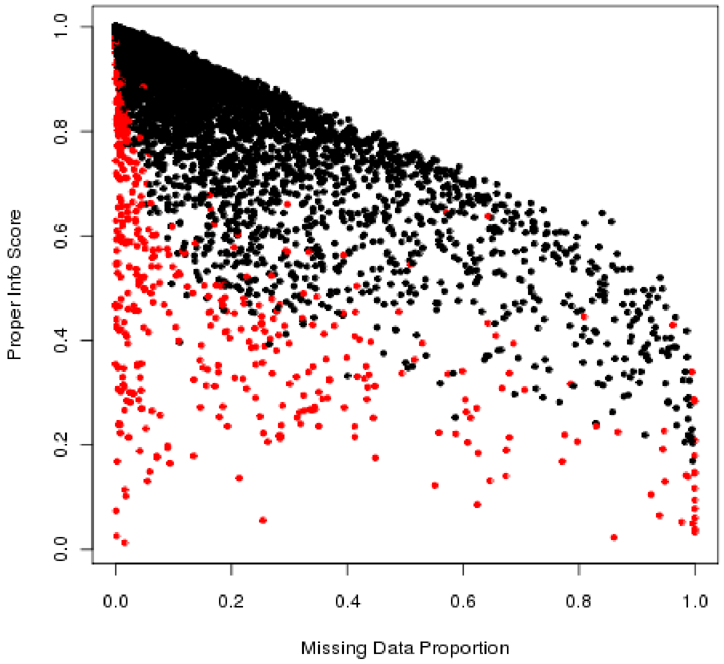
4.3.1 Final inclusion criteria for the imputed SNPs

One solution was to include only SNPs that are replicated in other datasets and then meta-analyse the data. This should remove SNPs that are significant solely because of a difference between imputed and genotyped methods. However, the method I employed, which entails ignoring the extra genotyped data in the controls, was to impute both the VQ58 cases and controls for the same SNPs using only the genotyped SNPs on the Hap300 array.

For imputed SNPs, I adopted inclusion criteria of an information score greater than 0.5, MAF greater than 0.01, HWE P value greater than 1×10^{-6} in cases and controls and a maximum genotype probability more than 0.9 for at least 5% of samples (chosen to equal the 95% call rate used for genotyped SNPs). The MAF, HWE and 'proper_info' score thresholds chosen for this study are consistent with previously published analyses of imputed data (Zeggini *et al.* 2008; Liu *et al.* 2010). I also filtered by maximum genotype probability as I noticed that when the information score is 0.5, there are SNPs where 80% of samples have a maximum genotype probability of less than 0.9 (see Figure 4.3). The analysis of these SNPs would be based on a dataset where just 20% of the samples had a high genotype certainty and indicates that simply filtering by information score might not be sufficient when dealing with imputed data.

Figure 4.3 The proportion of un-called SNPs plotted against the info score

The number of un-called SNPs on chromosome 9, calculated from the number of samples with a maximum genotype probability below 0.9, plotted against the proper information score. This plot shows that even if the information score is greater than 0.5 there can be SNPs with a high number of uncalled SNPs. The SNPs were imputed using IMPUTEv1 to the HapMap2 SNPs and shows a number of SNPs with a MAF less than 0.01 (red points), which have an information score greater than 0.5 and a very low number of uncalled genotypes. In most cases, this is because all the genotypes have been called with high certainty as the major homozygote, which is the most probable genotype if the minor allele is poorly represented in the reference panel.

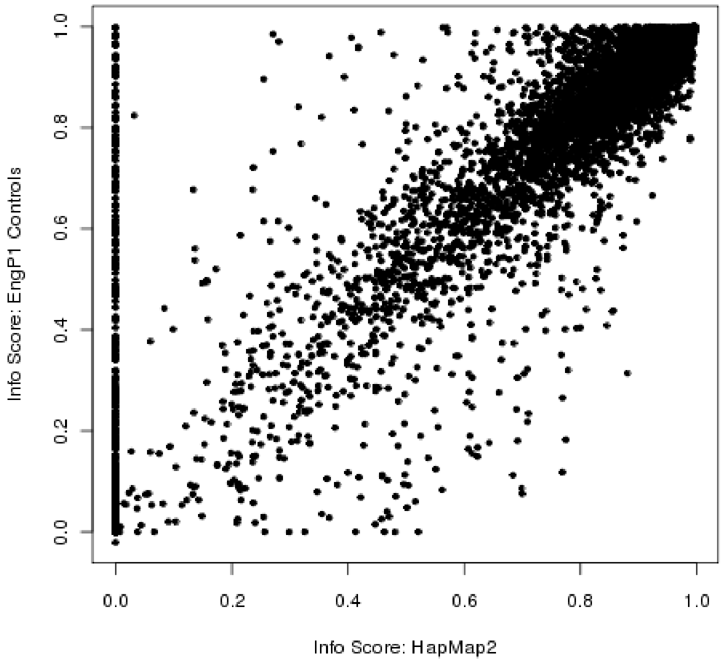


4.3.2 A comparison of reference panels used for imputation

I decided to compare the quality of imputation using different reference panels to determine which was the most appropriate, bearing in mind that VQ58 will impute with less confidence than the other datasets owing to being typed for a less dense panel of SNPs. In addition to imputation using IMPUTEv1 and the HapMap phase II reference panel, I also experimented with using IMPUTEv2 to impute the Hap550 SNPs with the EngP1 controls as a reference panel. I noticed from the initial imputation results that the information score for certain SNPs differs between these two methods (see Figure 4.4). The graph shows that there are a large number of SNPs with an information score of zero when using the HapMap II, 1,854 of which have an information score greater than 0.5 with the EngP1 reference panel. Although EngP1 would seem to be an optimal reference panel for imputing Hap550 SNPs, there was a subset of SNPs that achieved higher information scores with HapMap phase II. Of course, there are also SNPs that will not impute well with either reference panel as they are just not well tagged by the SNPs on the Hap300 array. These results suggested that I should use the breadth of HapMap phase II and the depth of the large number of samples in the EngP1 controls by using both reference panels together.

Figure 4.4 The information score for imputed SNPs with HapMap2 plotted against EngP1 as a reference panel

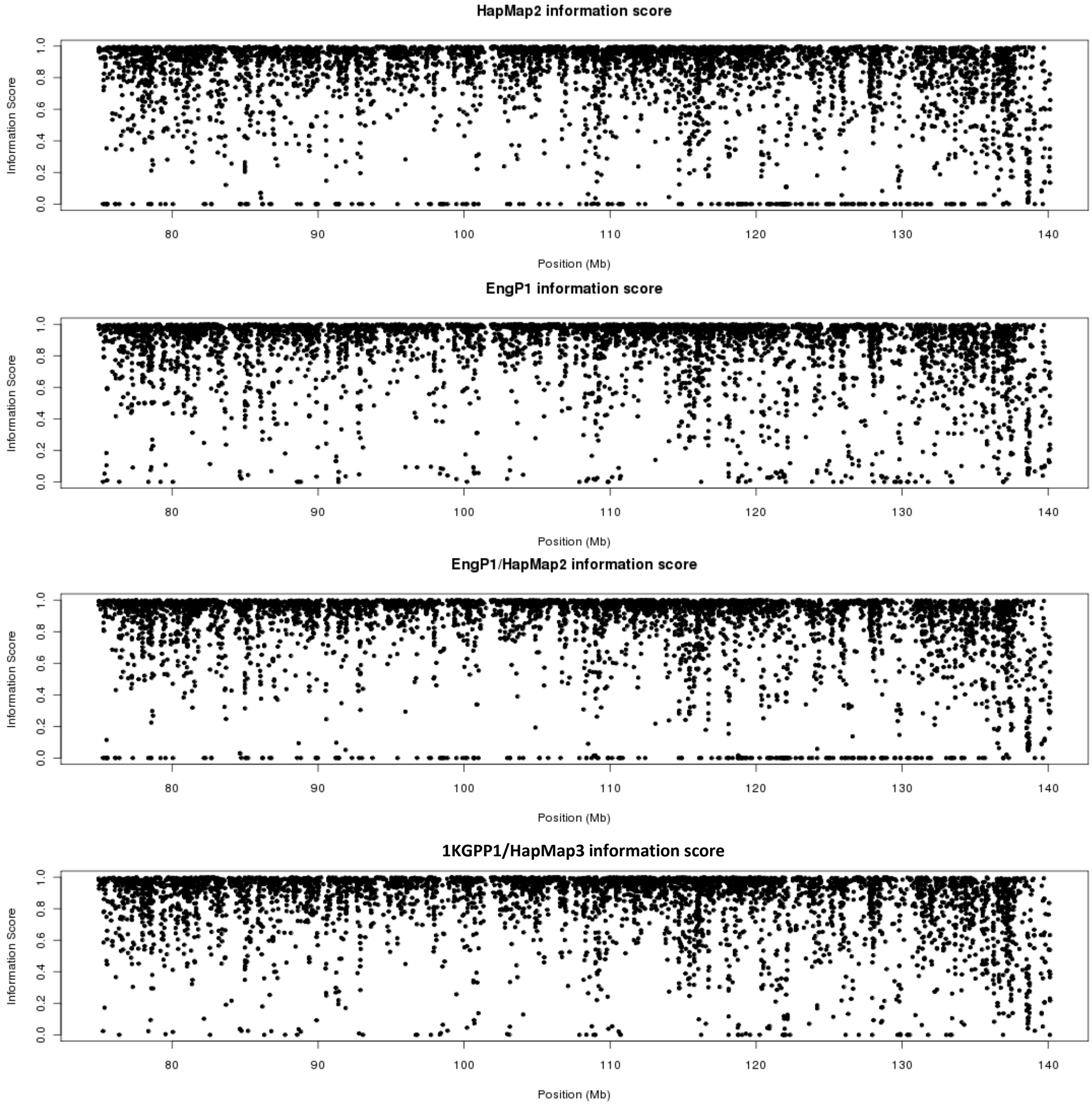
The plot shows the information score for imputing the same panel of SNPs in the VQ58 cases, using the HapMap2 or EngP1 reference panels. The correlation (r^2) of the information scores between the two methods is 0.931.



In an attempt to maximise the imputation quality and genotype certainty, I compared the information scores obtained using the different reference panel combinations available (see Figure 4.5) and the number of SNPs that would pass quality control criteria. I also imputed using the 1KGP pilot one data (1KGP P1) and HapMap3 reference panel to include in the comparison.

Figure 4.5 The IMPUTE information score comparison between different reference panels

Reference panels included are HapMap phase II, EngP1, EngP1/HapMap phase II, and the 1000 genomes project pilot one (1KGPP1)/HapMap phase3. A panel of 5328 SNPs from chromosome 9 were plotted for each reference panel used in the imputation to compare the information scores achieved with each method. From these plots it appears that the EngP1 and 1KGPP1/HM3 panels result in the least SNPs with an information score of zero. To ensure the comparison reflected the reference panel used and not the version of IMPUTE, I repeated the imputation to HapMap phase II using IMPUTEv2 for untyped SNPs on chromosome nine.



HapMap2 information score

EngP1 information score

EngP1/HapMap2 information score

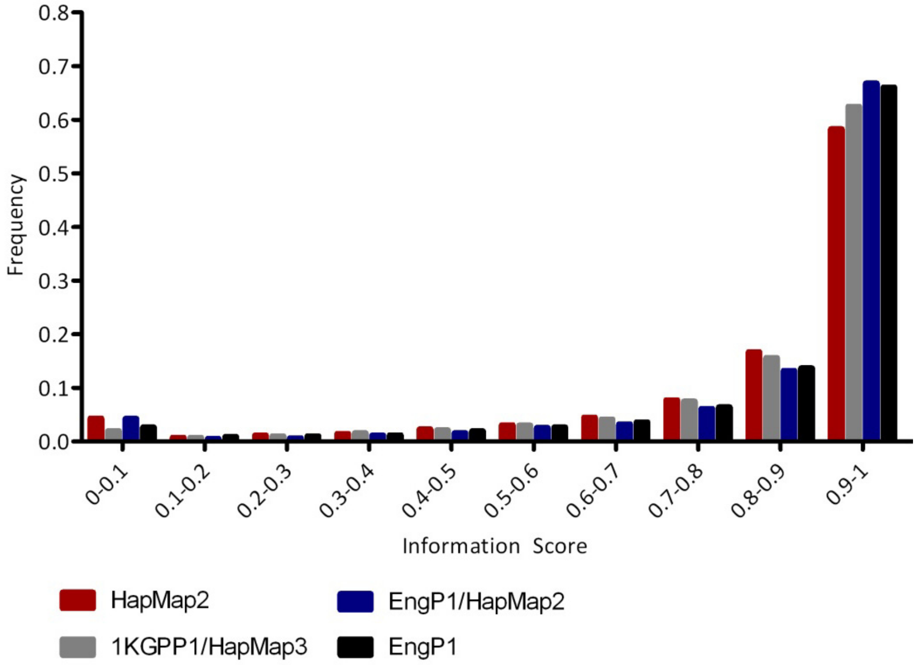
1KGPP1/HapMap3 information score

As HapMap II, 1KGPP1/HapMap3, and EngP1 consist of different numbers of SNPs, each of the plots in Figure 4.5 only include SNPs that were on the Hap550 array and present in every dataset. These plots illustrate that across the different reference panels, there are certain regions of the chromosome where SNPs are difficult to impute (owing to how well these SNPs are tagged by the genotyped SNPs) and that the distribution of SNPs with low information scores is very similar. The SNPs with information scores very close to zero generally had very low MAFs. Based on the information score alone, the EngP1 and 1KGP/HapMap3 reference panels achieve the lowest number of SNPs with an information score less than 0.1.

In Figure 4.6 (below), I have used the same data to compare the distribution of the information score obtained using each reference panel and plotted the results in a histogram. The reference panels do not vary greatly and the numbers of SNPs in each range are comparable between EngP1 and EngP1/HapMap2. The only range to show a large difference is 0.9-1, where EngP1/HapMap2 outperforms the other panels.

Figure 4.6 The distribution of info scores after imputation using different reference panels in VQ58.

The figure shows the distribution of the information score from IMPUTE with each of the reference panels for the SNPs on chromosome 9. The data is the same as that used in Figure 4.5 and the SNPs included are the same for each dataset. The number of SNPs for each dataset is comparable in each info score range, except in 0.9-1 where EngP1/HapMap2 seems to outperform the other reference panels.

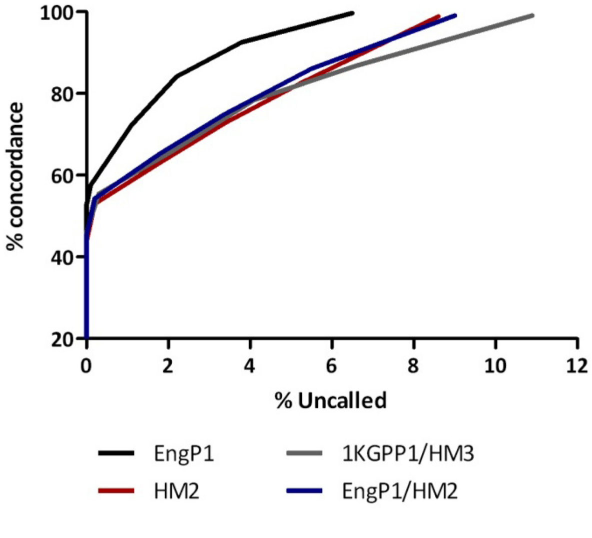


In terms of the number of SNPs with an information score less than 0.1, EngP1/HapMap phase II had 429, EngP1 had 266, HapMap phase II had 431 and 1KPG/HapMap3 had 194 SNPs. Each reference panel contained different numbers of samples, which could impact the quality of imputation for certain SNPs. EngP1 consisted of almost 928 samples, HapMap phase II contains 90 samples, and 1KGP pilot one and HapMap3 consist of 60 and 165 samples, respectively.

Further investigation of the 429 SNPs from the EngP1/HapMap phase II imputation to determine the MAF of the SNPs revealed that 425 had a MAF of zero. Therefore, even with the smaller number of samples compared to EngP1, the 1KGP/HapMap3 panel performs best over these SNPs. However, EngP1 produced the highest concordance rate when genotyped SNPs were imputed (see Figure 4.7), while the other reference panels performed similarly to each other. Overall, the concordance between imputed and genotyped results increased as the proportion of samples failing to meet the genotype calling threshold increased.

Figure 4.7 Concordance of imputed SNP genotypes vs proportion uncalled genotypes

The data for this figure are from a 7Mb segment of chromosome 9 that contained 761 genotyped SNPs. The concordance is calculated by imputing the genotyped SNPs by leaving one out at a time at performing the imputation based on the remaining SNPs and then comparing the results. Although it is not a good practise to convert probabilities into 'best guess' genotypes for analysis, it is adequate for a comparison of the concordance of imputed SNPs against genotyped SNPs. The percentage uncalled is determined by the number of genotypes with a maximum genotype probability less than 0.9. As this threshold increases the concordance decreases.



The method to predict the missing Hap550 SNPs in the VQ58 dataset was largely decided based on the number of imputed Hap550 SNPs that passed the QC criteria. The results are given in Table 4.1. The imputation method that resulted in the highest number of successfully imputed Hap550 SNPs (112,986 SNPs) used the EngP1/HapMap reference panels.

Table 4.1 A summary of imputation performance in VQ58 using different reference panels

The data in this table relates to SNPs in the VQ58 cases. The imputation using just the HapMap2 reference panel only included those SNPs on the Hap550 array. The number of Hap550 array SNPs in the EngP1 panel is less than that of the HapMap2 panel because only 491,199 SNPs passed the quality control criteria in EngP1 after removing SNPs that failed genotyping or were out of HWE ($P<1x10^{-6}$). The figure in brackets is the percentage of Hap550 SNPs out of the full 239,855 Hap550 SNPs that were not included in the Hap300 array. The bottom row illustrates the increased number of successfully imputed SNPs if the samples are genotyped for the Hap550 SNPs compared to the Hap300.

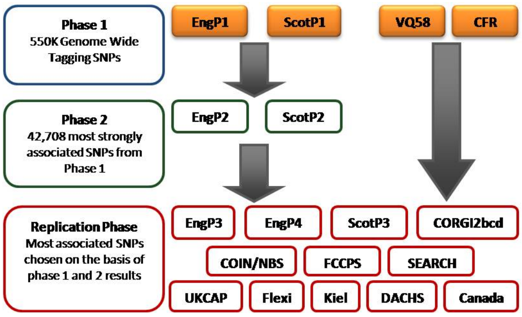
SNP genotyping platform	Reference panel	No. of SNPs in output	No. of Hap550 SNPs imputed	SNPs passing QC criteria		
				Total No. SNPs	No. imputed Hap550 SNPs	% imputed Hap550 SNPs
Hap300	HapMap2	531,356	239,855	396,811	105,336	43.9
Hap300	EngP1	491,199	198,720	397,622	105,177	52.9 (43.8)
Hap300	EngP1/ Hapmap2	3,842,830	239,855	1,721,574	113,719	47.4
Hap550	HapMap2	3,842,830	-	2,129,771	-	-

After the quality control checks in cases and controls, 292,208 SNPs that were genotyped in VQ58 were also included in the HapMap2 reference panel and hence used for the imputation. There were 292,479 SNPs that were genotyped in VQ58 and included in the EngP1 reference panel. The number of SNPs successfully imputed was comparable to the findings of Anderson *et al.*, who imputed a similar dataset using HapMap phase II (Anderson *et al.* 2008). The optimal method for imputing Hap550 SNPs from Hap300 genotypes is, thus, to use the EngP1 and HapMap phase II reference panels together, as this led to an additional 8,542 SNPs compared to using

EngP1 alone. The imputed data was incorporated into the meta-analysis described in Section 3.7 in the previous Chapter.

4.4 Imputation and meta-analysis of HapMap phase II SNPs

Figure 4.8 The datasets included in the meta-analysis of imputed HapMap2 SNPs


Phase 1
550K Genome Wide Tagging SNPs
EngP1
ScotP1
VQ58
CFR
Phase 2
42,708 most strongly associated SNPs from Phase 1
EngP2
ScotP2
Replication Phase
Most associated SNPs chosen on the basis of phase 1 and 2 results
EngP3
EngP4
ScotP3
CORGI2bcd
COIN/NBS
FCCPS
SEARCH
UKCAP
Flexi
Kiel
DACHS
Canada

The aim of this analysis is to attempt to refine the associated loci that have already been discovered by identifying additional SNPs through meta-analysis of the imputed SNPs. In addition to the EngP1, ScotP1 and VQ58 datasets already studied, I have also included an additional case control dataset, CFR, which was genotyped on the Illumina 1M SNP array.

I have performed this analysis using the same quality control criteria as for the imputation of VQ58, where SNPs were rejected if the information score was below 0.5, the proportion of missing data less than 5% and MAF greater than 0.01. Instead of using threshold genotypes for the analysis, the ‘proper’ method uses a test based on the missing data likelihood, which fully takes into account the uncertainty of the genotypes. The P value used in the following analyses is the frequentist additive score test, which is the Cochran-Armitage test for additive genetic effects.

For this analysis all datasets (see Figure 4.8) were imputed to the HapMap phase II reference panel, except VQ58 which was imputed to EngP1/HapMap phase II to attempt to maximise the number of Hap550 SNPs in the analysis. The HapMap3/IKGP P1 reference panels were not used for this analysis as it was considered that the low number of samples in the pilot data and the time and computational cost to repeat the imputation did not offer a substantial advantage over HapMap phase II for common variants. Equally, the 1KGP P1 data did not contain enough samples to allow adequate representation of low frequency alleles and rare variants, where ideally hundreds more samples will be needed to provide the additional power required. The use of this reference panel is explored later in this Chapter.

The number of SNPs that pass the inclusion threshold was 2,129,771 in EngP1, 2,130,884 in ScotP1, 2,259,118 in CFR and 1,721,575 in VQ58. The number of SNPs overlapping all groups was 1,678,943 SNPs and I, therefore, also included SNPs that were not present in VQ58 into the analysis.

4.4.1 Results of the meta-analysis after imputation

The association results from the meta-analysis of the four GWA study datasets are summarised in the Manhattan plot below (Figure 4.9). Regions containing SNPs with P values less than 1x10^{-6} were plotted individually to gain information about the LD between SNPs and a closer view of the region. There were suggestive regions on chromosome 8, 10, 11, 12, 15, 18, but we have identified significant SNPs in these regions previously (see Figure 4.10 to 4.15 below). However, the results do highlight many SNPs with lower P values than those previously identified that may aid the fine-mapping of these regions.

Figure 4.9 Manhattan Plots for imputed and genotyped SNPs

A) The Manhattan plot summarising the results of the meta analysis of EngP1, VQ58, ScotP1 and CFR for the genotyped and imputed HapMap SNPs. B) the same analysis showing just the genotyped SNPs

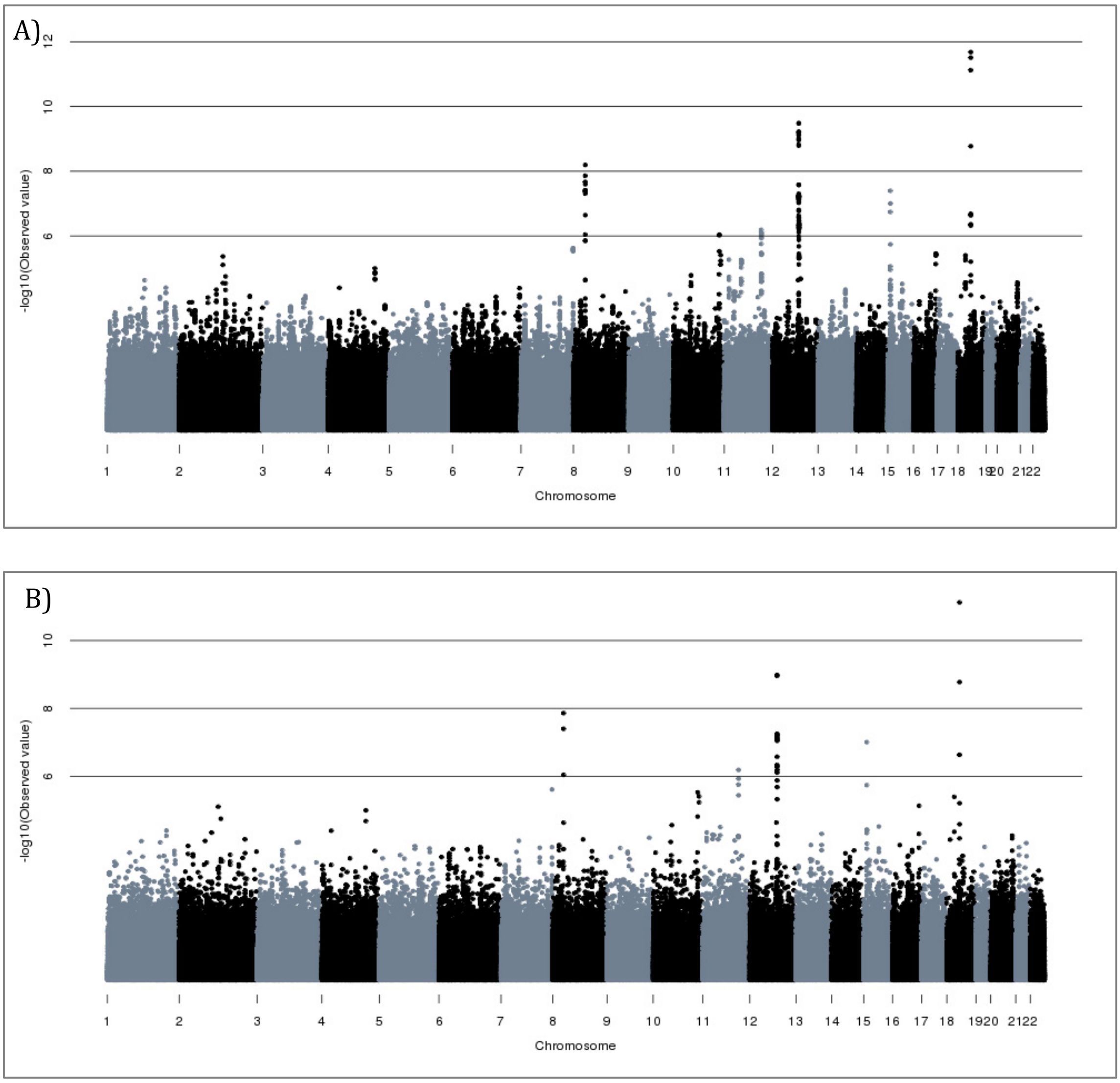


Figure 4.10 The significant SNP region on chromosome 8

The -log P values plotted against location. The SNPs plotted as diamonds are genotyped, while triangles are imputed SNPs. The large diamond is the original top SNP, rs6983267, from the GWA study for EngP1 and LD is in relation to this SNP.

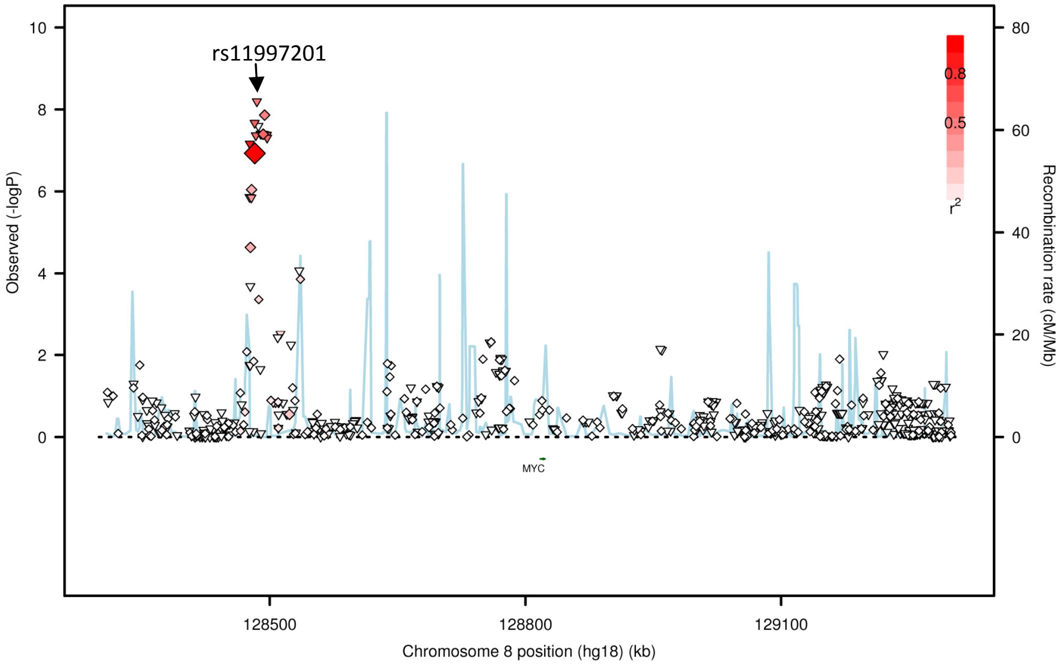


Figure 4.11 The association results for HapMap2 SNPs on chr10

The SNPs plotted as diamonds are genotyped, while triangles are imputed SNPs. The values for pairwise LD by r^2 were calculated in relation to rs10795668 (large diamond) using HapMap CEU. The SNPs plotted as diamonds are genotyped, while triangles are imputed SNPs.

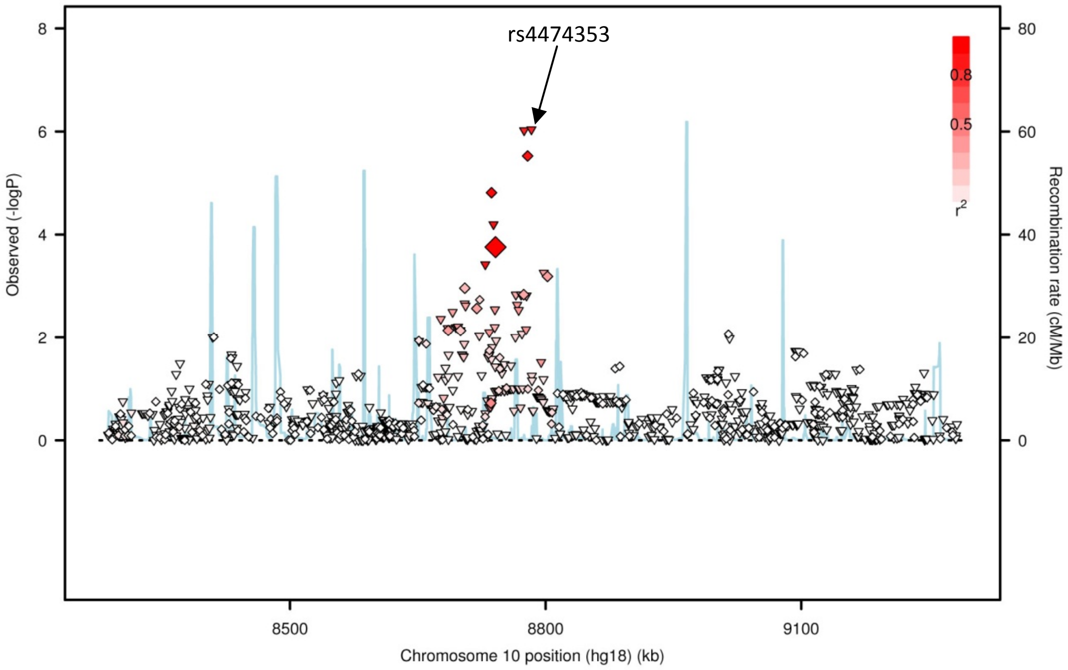
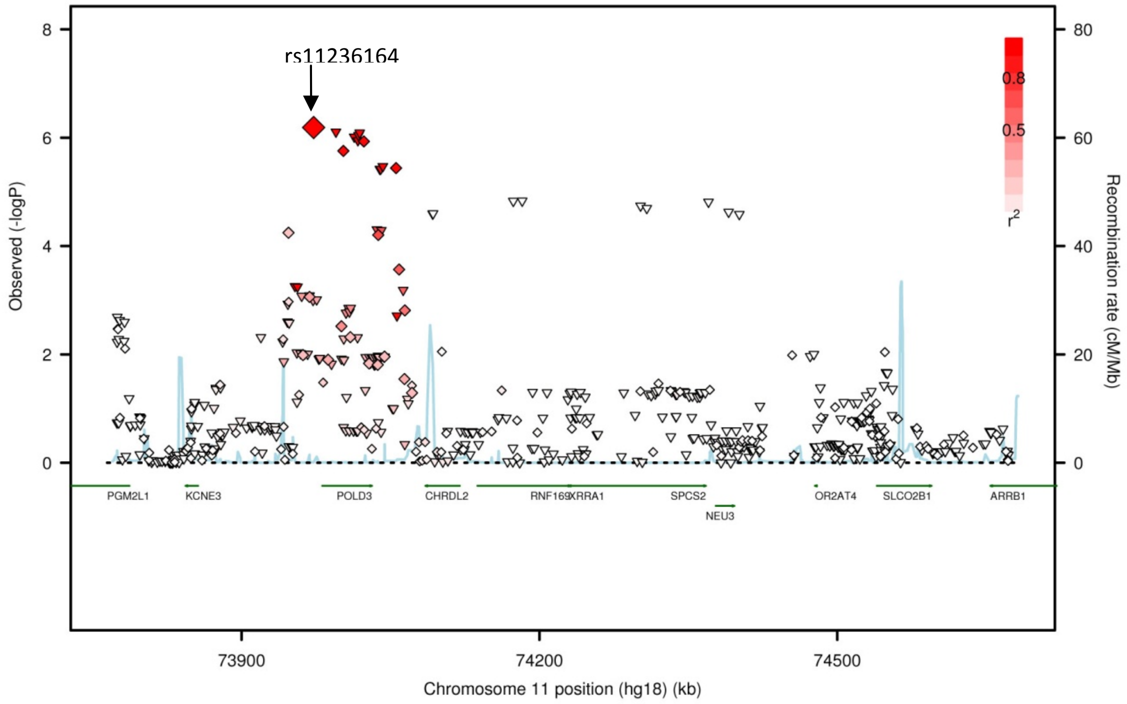


Figure 4.12 The association results for HapMap2 SNPs on chromosome 11

The SNPs plotted as diamonds are genotyped, while triangles are imputed SNPs. Pairwise LD by r^2 was calculated in relation to rs11236164 (this SNP was genotyped)



The most strongly associated SNP for chromosome 12 was rs7972465, which was imputed and is located at 48,832,393bp ($P=3.27x10^{-7}$, beta=-0.19, se=0.031, see Figure 4.13, below). This SNP is located within the gene, LAG1 homolog of ceramide synthase 5 (LASS5). However, if EngP2 and ScotP2 are included (and CFR excluded) the P value for this SNP is $5.07x10^{-5}$ suggesting that the association is not replicated in these datasets. These plots show that the likely location of the causal variant is difficult to pin down. In the top plot, the pairwise r^2, indicated by the shading, is in relation to rs11169552. The bottom plot shows the area redrawn with LD calculated in relation to rs7136702 and it is clear that the more strongly associated SNPs to the left of the plot are actually in high r^2 with this SNP. However, LD extends to encompass several likely genes.

Figure 4.13 The association results for HapMap2 SNPs on chromosome 12

The SNPs plotted as diamonds are genotyped, while triangles are imputed. The two plots below cover the same region and same SNPs on chromosome 12. The top shows LD in relation to rs11169552 (large red diamond) and the bottom plot the LD in relation to rs7136702 (large red diamond). The most significant SNPs are all to be in moderate to high LD with rs7136702.

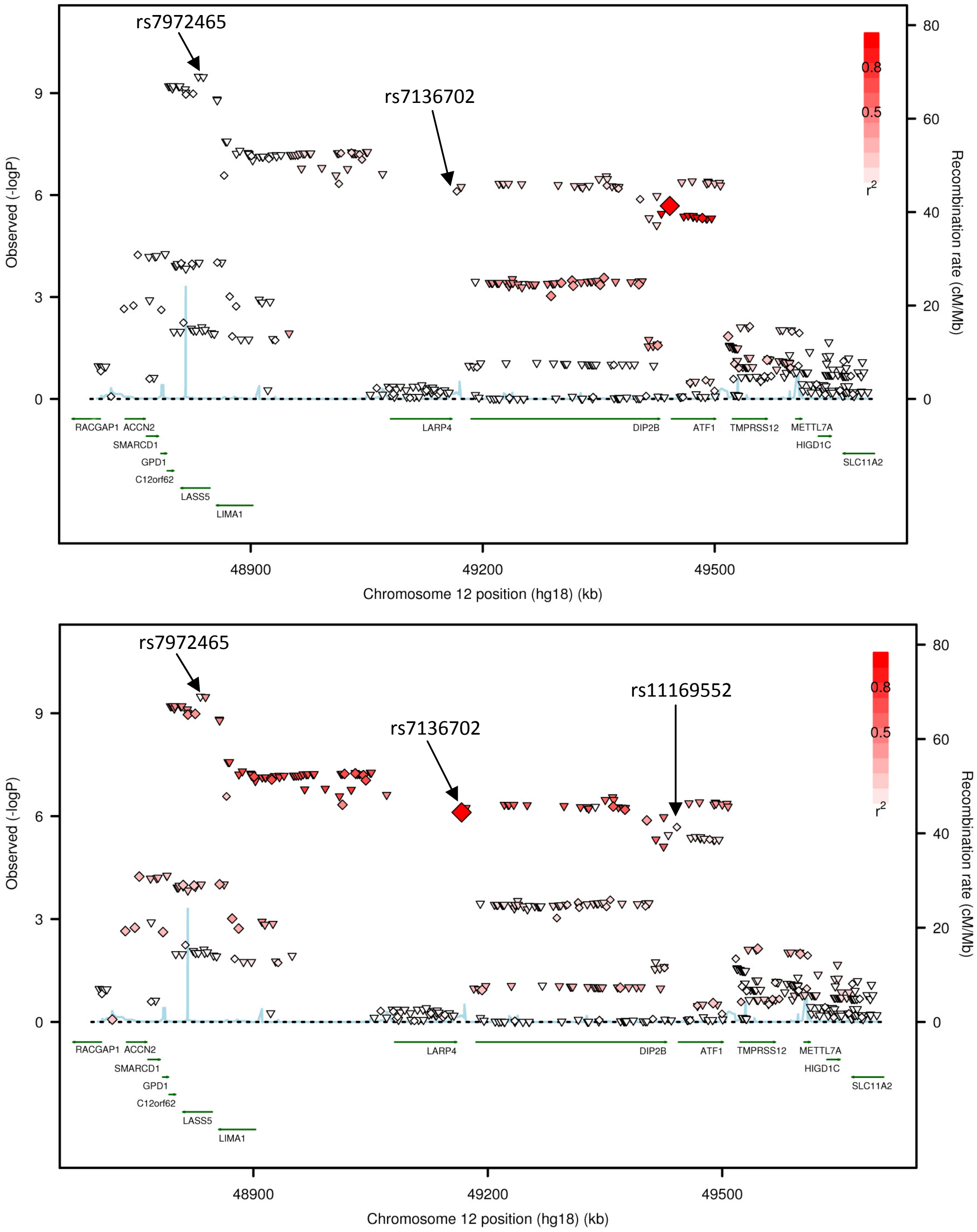


Figure 4.14 The association results for the HapMap2 SNPs on chromosome 15

The pairwise LD was calculated in relation to rs4779584 using the HapMap2 CEU dataset. The SNPs plotted as diamonds are genotyped, while triangles are imputed SNPs.

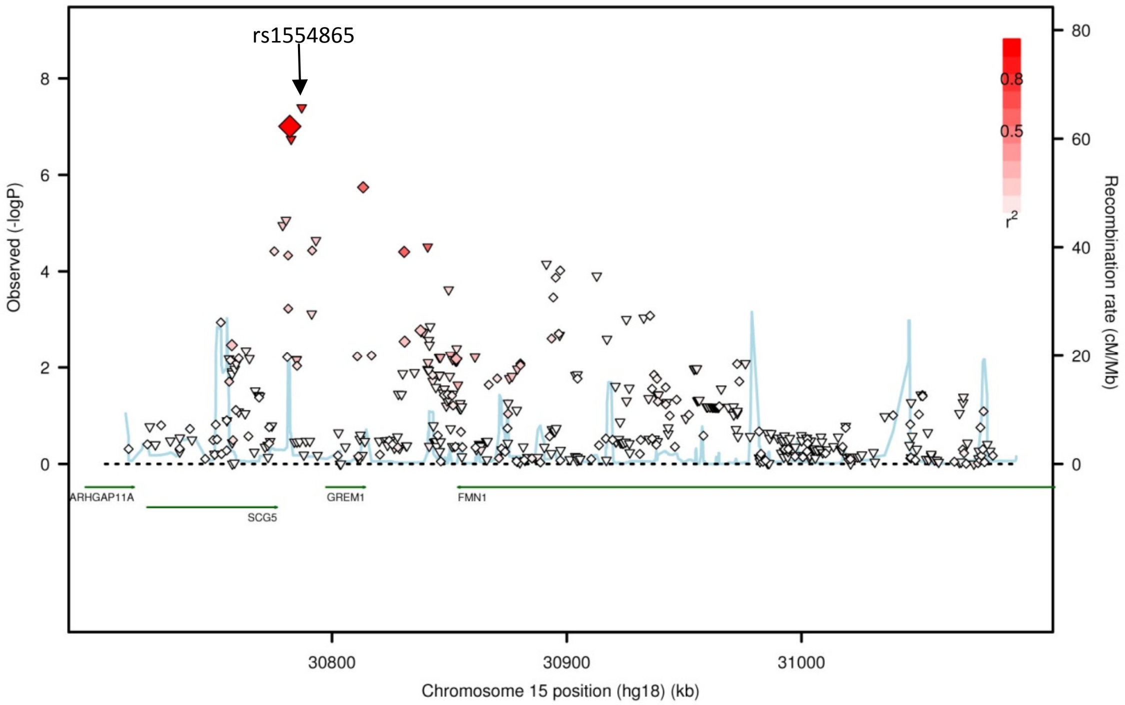
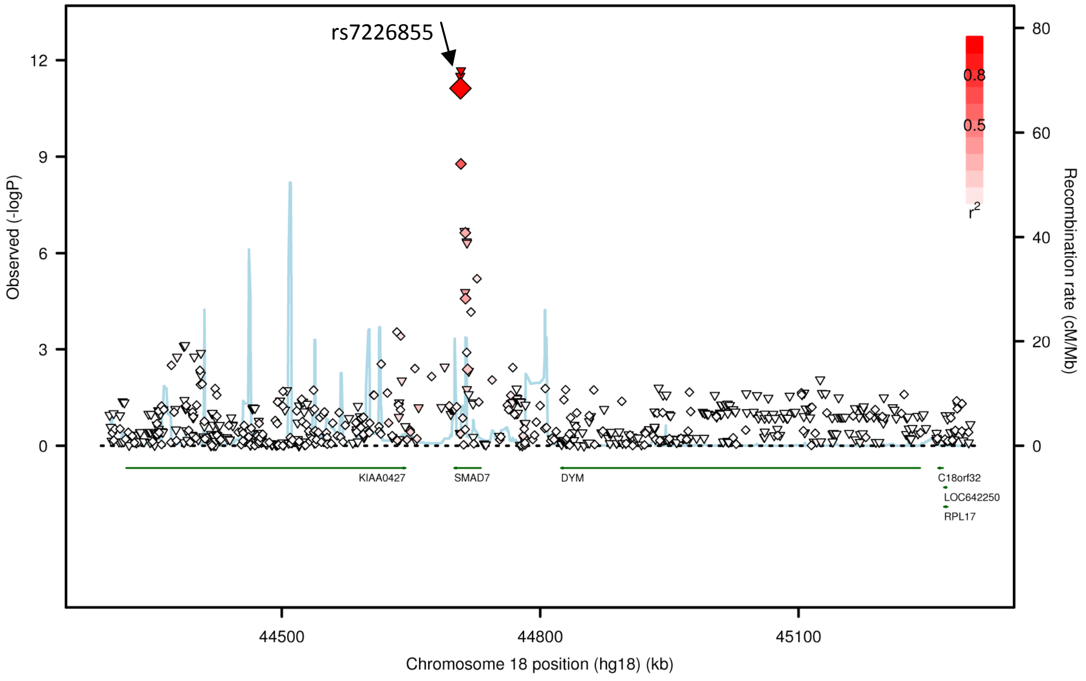


Figure 4.15 The association for the HapMap2 SNPs on chromosome 18

The pairwise LD was calculated in relation to rs4939827 using the HapMap2 CEU dataset. The SNPs plotted as diamonds are genotyped, while triangles are imputed SNPs.



The results of the analysis with HapMap phase II imputed SNPs show that for most of the identified loci, with the exception of chromosome 11, the imputed SNPs are more strongly associated with disease than those that were genotyped in the study (See Table 4.2).

Table 4.2 The most associated SNPs after imputation to HapMap2 and meta-analysis in four GWA study datasets (EngP1, ScotP1, VQ58 and CFR).

The SNP positions are from the Human Genome build 36. Alleles are coded A/B and beta is with reference to allele B. P for between-study heterogeneity was greater than 0.05 for all SNPs included.

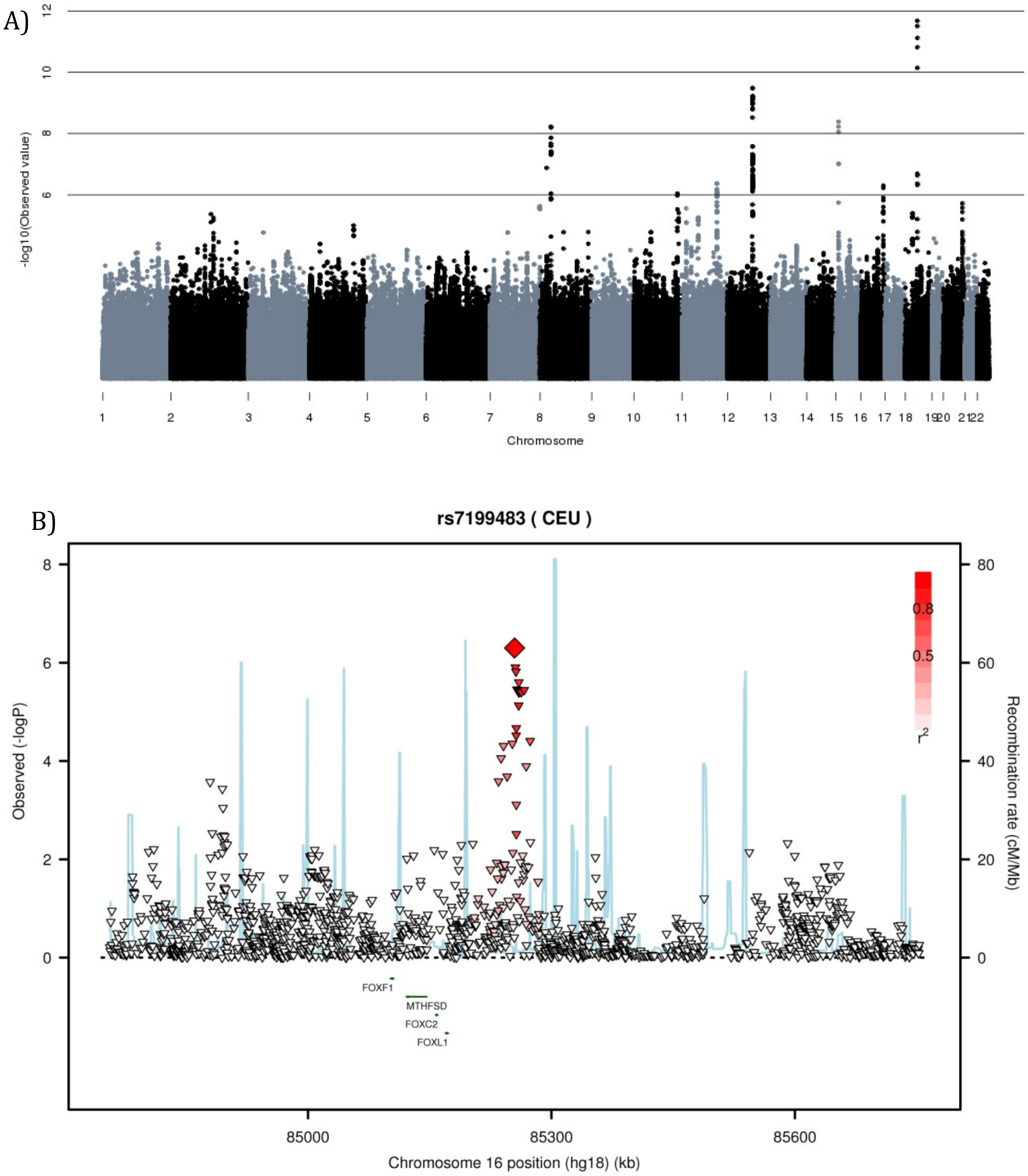
Chr.	SNP	Position (bp)	Alleles (A/B)	P value (allelic)	Beta (logOR)	SE	Type
8	rs11997201	128484916	A/C	$6.39x10^{-9}$	-0.147	0.030	I
10	rs4474353	8783319	A/G	$9.14x10^{-7}$	0.157	0.032	I
11	rs11236164	73972614	A/C	$6.49x10^{-7}$	-0.147	0.030	G
12	rs7972465	48832392	G/T	$3.27x10^{-10}$	-0.194	0.031	I
15	rs1554865	30787098	C/T	$4.02x10^{-8}$	-0.251	0.046	I
18	rs7226855	44708046	A/G	$2.08x10^{-12}$	-0.208	0.030	I

4.4.2 Meta-analysis of imputed SNPs without filtering by missing proportion

Incidentally, as it is quite stringent to reject SNPs on the basis of the number of samples with a maximum probability less than 0.9, considering that I have analysed the genotypes using the ‘proper’ frequentist methods, instead of best guess genotypes, I also analysed the data without filtering on this aspect. If the missing proportion is not taken into consideration, 2,329,091 SNPs pass the inclusion threshold. However, the results identified just one additional peak on chromosome 16 with a P value less than $1x10^{-6}$ (see Figure 4.16). The most strongly associated SNP is rs7199483 ($P=5.02x10^{-7}$, beta=0.185 (with reference to the T allele), SE=0.037, 85,254,509bp). This region contains 14 imputed SNPs with P values less than $1x10^{-5}$ and is not the same region as that of the previously discovered SNP rs9929218, which was located at 67,378,447bp. The SNP rs7199483 is located within a spliced EST BQ774482, and is in a region containing enhancer promoter Histone marks (H3K4Me1), suggesting that it could be involved in regulation. This region has since been identified in a new meta-analysis using genotyped SNPs and is being followed up.

Figure 4.16 Meta-analysis of imputed SNPs without missing proportion filter

A) The genome wide association results and B) the associated SNPs on the chr16 region not identified in the previous analysis. The imputed SNPs are plotted as a triangle, genotyped as a diamond. The large red diamond is rs7199483, which is imputed, and indicates that LD is in relation to this SNP.



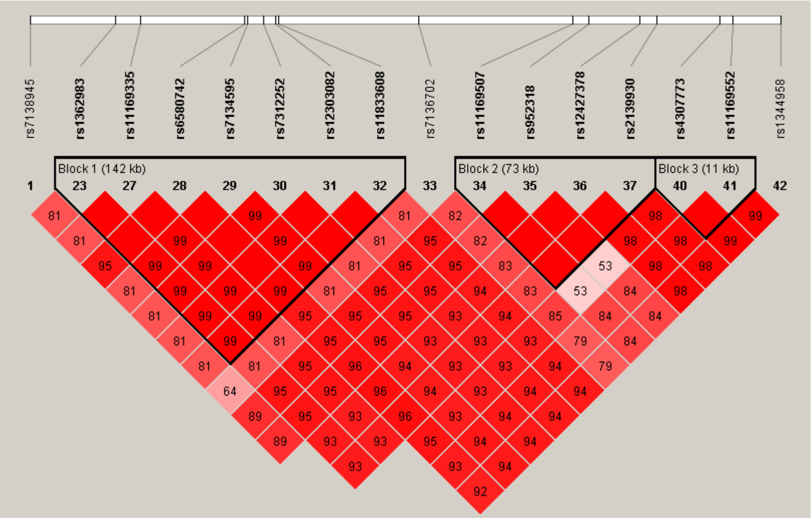
4.5 Further analysis of the CRC associated SNPs

4.5.1 Are SNPs identified in the same locus really independent?

There were several loci identified (described in Chapter 3, section 3.7), where two SNPs were significantly associated with CRC. To determine if these SNPs are truly independent, I performed some additional analyses, including a logistic regression analysis incorporating genotypes from 45,130 samples. This analysis included three additional replication case/control datasets that incorporate population controls, LiLi (Kentucky), Pavel (Prague) and EPICOLON.

The two SNPs rs11169552 and rs7136702 are not in high LD (r^2=0.12, see Figure 3.16, Chapter 3), however, pair-wise D' between them is 0.79. In fact, all of the SNPs identified on chromosome 12 are located within an 'LD block' (see Figure 4.17, below). To determine if these two SNPs were truly independent, I performed a logistic regression analysis. An association analysis using an unconditional logistic model for the two SNPs gave P values of $3.78\text{x}10^{-10}$ for rs11169552 (OR=1.10) and $1.67\text{x}10^{-7}$ for rs7136702 (OR=1.08). When the results were conditioned on rs7136702, the P value for rs11169552 increased to $6.50\text{x}10^{-7}$ (OR=1.09). However, it is difficult to say from these results whether the effects of the SNPs are independent.

Figure 4.17 The LD by D' for the chromosome 12 SNPs



The pair of SNPs on chromosome 1, rs6691170 and rs6687758, are also in low r^2 (0.15), while pair-wise D' is 0.64 (see Figure 3.16, in Chapter 3). The association analysis under an unconditional logistic model gave P values of $4.25x10^{-10}$ (OR=1.09) and $3.99x10^{-10}$ (OR=1.11), respectively. However, the analysis for rs6687758 gave a P value of $2.48x10^{-4}$ (OR=1.07) when conditioned on the effect of rs6691170. All ORs were calculated with reference to the risk allele. The results for these SNPs also indicated that they were not independent of one another.

To assess the overall risk of the high risk haplotype for each of these pairs of SNPs, I performed a haplotype analysis (using 'hap-logistic' in PLINK). The increased risk associated with each of these pairs is modest, but the results showed that individuals with the high risk haplotype of TC for rs7136702 and rs11169552 had an increased risk of 1.09 fold ($P=8.92x10^{-9}$, frequency=0.34) compared to the low risk haplotype (see Table 4.3). The same analysis for rs6691170 and rs6687758 showed that an individual with the high risk haplotype, TG, had an increased risk of 1.15 fold ($P=1.51x10^{-13}$, frequency=0.20) compared to the low risk haplotype. The results for both pairs of SNPs showed that carrying the risk allele for both SNPs is significantly associated with disease risk, while carrying the risk allele for just one of the SNPs is not. However, for rs7136702 and rs11169552, the results for the TT haplotype show some evidence of a protective effect with an OR of 0.899, with one low risk allele at rs11169552. These analyses provide evidence that the effects of the two SNPs identified at each locus are not independent.

Table 4.3 Analysis of haplotype risk for the pairs of chr1 and chr12 SNPs

The frequencies and disease risk associated with the two pairs of significant SNPs identified over each possible haplotypes. The low and high risk haplotypes for significantly associated for both, but it appears that both risk alleles are required.

SNPs	Haplotype	Risk	Freq (aff)	Freq (ctrl)	P value	OR
rs7136702, rs11169552	TT	H/L	0.021	0.022	0.069	0.899
	CT	L/L	0.233	0.251	$3.62x10-^{10}$	0.903
	TC	H/H	0.348	0.329	$8.92x10^{-9}$	1.09
	CC	L/H	0.398	0.397	0.733	1
rs6691170, rs6687758	TG	H/H	0.176	0.158	$1.51x10^{-13}$	1.15
	GG	L/H	0.037	0.038	0.289	0.959
	TA	H/L	0.203	0.200	0.353	1.02
	GA	L/L	0.584	0.604	$3.12x10^{-9}$	0.92

4.5.2 Epistasis analysis of the 16 identified CRC susceptibility SNPs

Incorporating just the cases, I performed a pair-wise gene-gene interaction analysis using the epistasis command in PLINK on the 16 discovered SNPs, rs6983267 (8q24), rs4939827 (18q21), rs4444235 (14q22.2), rs4779584 (15q13), rs16892766 (8q23.3), rs10795668 (10p14), rs3802842 (11q23.1), rs9929218 (16q22.1), rs10411210 (19q13.1), rs961253 (20p12.3), rs6691170 (1q41), rs6687758 (1q41), rs10936599 (3q26), rs7136702 (12q13), rs11169552 (12q13) and rs4925386 (20q13). Only one pair of SNPs, rs6687758 and rs7136702, showed suggestive evidence of epistasis

($P=9.61\times10^{-6}$, OR=0.90). This evidence was maintained after correction for multiple testing (P= 0.0015).

4.5.3 Imputation and analysis of 1KPG/HapMap3 SNPs for four loci

In order to investigate the use of the 1KGPP1/HapMap3 reference panel and compare the association results with imputation using HapMap Phase II, I imputed SNPs in the region of four associated loci (on chr12 (rs7136702 and rs11169552), chr1 (rs6691170 and rs6687758), chr3 (rs10936599) and chr20 (rs4925386), described in Section 3.7). Imputing using this reference panel also allowed us some insight into how many additional SNPs would be detected and imputed successfully. Although, the depth of the 1KGP P1 data was small, the increase in common variants in comparison to HapMap phase II made it worthwhile to attempt to use this data to help fine-map the association signal at these loci. I also attempted to impute EngP2 and ScotP2, where despite the lower overall density of genotyped SNPs a reasonable number of untyped SNPs were successfully imputed (see Table 4.4).

Table 4.4 A summary of imputed SNPs from four associated loci

This table shows the region that was imputed using the 1KGPP1 reference panel and the number of SNPs that passed quality control for the GWA datasets, EngP1, ScotP1 and VQ58 and also the number in EngP2 and ScotP2 after imputation.

Chr.	Region to impute		Total SNPs passing quality control			
	Start (kb)	End (kb)	Total Genotyped	Total after imputation	Total genotyped EngP2 and ScotP2	Total after imputation EngP2 and ScotP2
1	220,000	221,000	76	630	17	126
3	170,000	172,000	260	2199	27	258
12	48,000	50,000	158	2736	57	956
20	60,000	61,000	129	946	20	139

The results of the meta-analysis (using SNPTEST and META) of all imputed SNPs (including those in EngP2 and ScotP2 wherever possible) and the location of genes in the region are summarised in the figure below (see Figure 4.18). The region on chromosome 12 had the highest number of SNPs and it is clear from the plots that there are many imputed SNPs with stronger association signals than those that were genotyped (see Figure 4.18 C). Indeed, the results showed 222 SNPs with lower P values (all present in EngP2 and ScotP2). The most strongly associated SNP was rs12582180 at 49,053,552bp (P=1.91x10^{-6}, beta=-0.11 (se=0.023)). This SNP was not included in the HapMap phase II imputation described above and is in LD (r^{2}) with rs7136702 and located within a provisional gene of unknown function, *FAM1864*, and in close proximity (downstream) to the *LARP4* (La ribonucleoprotein domain family member 4) gene.

The original genotyped SNP, rs7136702, is located upstream of LARP4, while rs11169552 is located between activating transcription factor 1 (ATF1) and disco interacting protein 2B (DIP2B), which may have a function in epithelial cell determination.

In this analysis, the locus on chromosome 3 is the only one where the signal is not refined by imputation. The strongest imputed SNP rs35446936 at 170,969,202bp (P=2.88x10^{-6}, beta =-0.121 (se=0.026)), while the most strongly associated genotyped SNP is this region is rs10936599 (discussed previously), which lies within the coding region of the myoneurin gene (*MYNN*), which encodes a zinc finger domain containing protein involved in the control of gene expression. Other nearby genes include the telomerase gene (*TERC*) and the actin-related protein M1 (*ARPM1*). The opposite (minor) allele of rs10936599 is associated with increased risk of Celiac disease (Dubois *et al.* 2010).

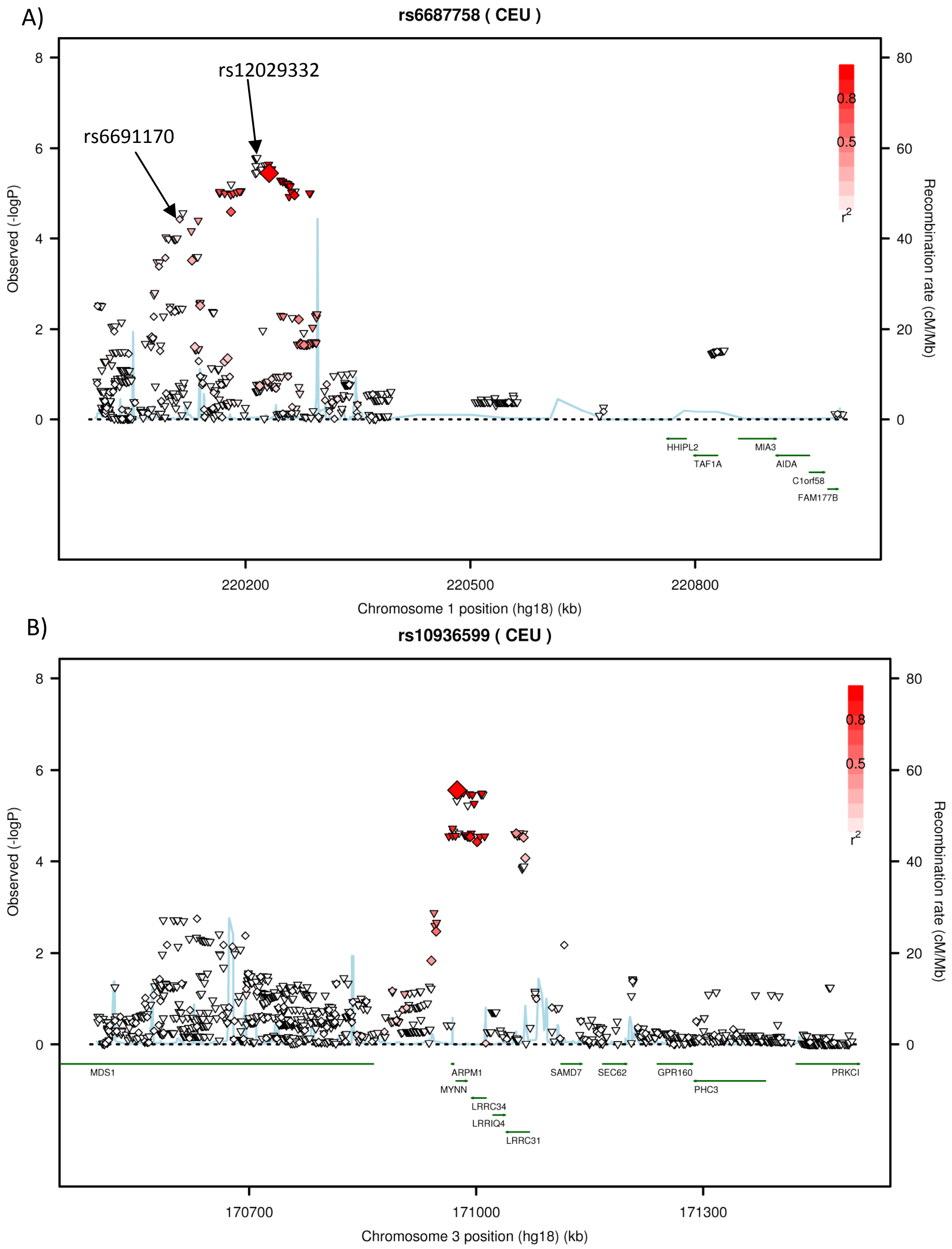
There were 45 imputed SNPs with P values better than or equal to the two genotyped SNPs on chromosome 1, with the most strongly associated being rs12029332 at 220,215,445bp (P=1.62x10^{-6}, beta=0.13 (se=0.03). This SNP was successfully imputed in the phase two datasets and is 16Kb downstream of the genotyped SNP rs6687758. rs12029332 is located in a region that contains a number of spliced ESTs and is 238Kb upstream of the gene dual specificity phosphatase 10 isoform (DUSP10, located at 219,941,389-219,977,425bp), which is involved in the negative regulation of MAPK/ERK, p38 and SAPK/JNK.

The results for chromosome 20 highlight just one imputed SNP that was more strongly associated with CRC, rs624313 located at 60,360,807bp (P=5.94x10^{-6}, beta=0.13 (se=0.029)). Both this SNP and the genotyped SNP, rs4925386, are situated within the Laminin alpha 5 (LAMA5) gene (found at 60,317,516-60,375,763bp). Laminins are thought to regulate the attachment, migration and organisation of cells to form tissues during embryonic development and is thought to induce the expression of noggin, a secreted BMP antagonist. This analysis did not include EngP2, as this SNP did not pass the inclusion criteria in this dataset.

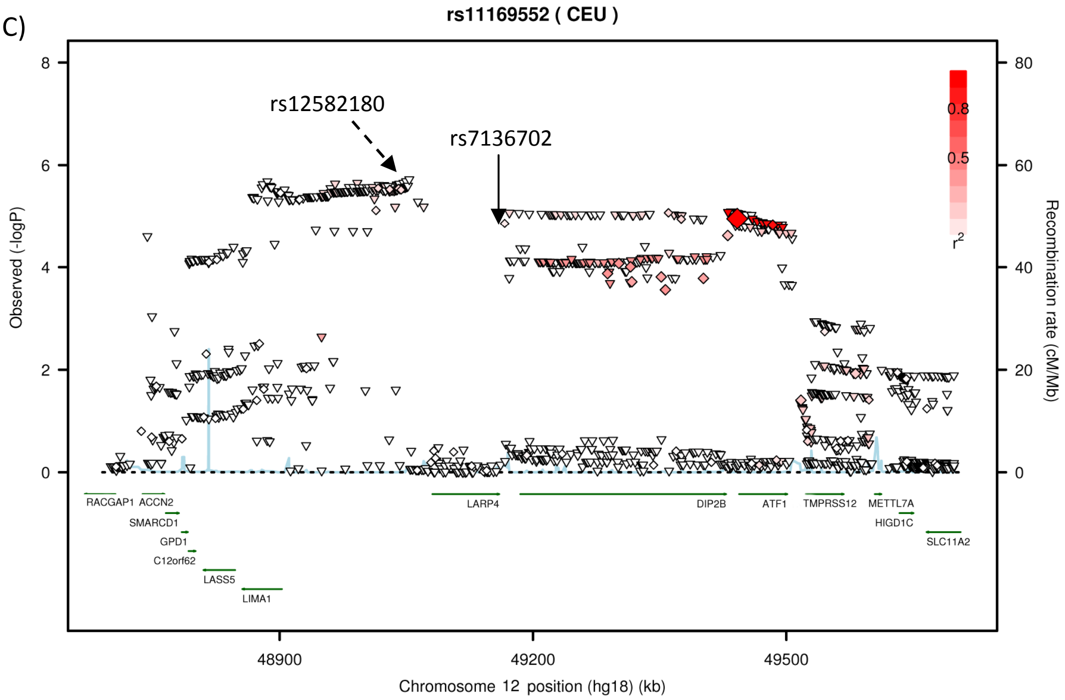
These results require further analysis to confirm whether these more strongly associated SNPs tag the causal variant better than those that were genotyped.

Figure 4.18 The results of the meta analysis of the four loci to include imputed SNPs in the 1KGP/HapMap3 panel

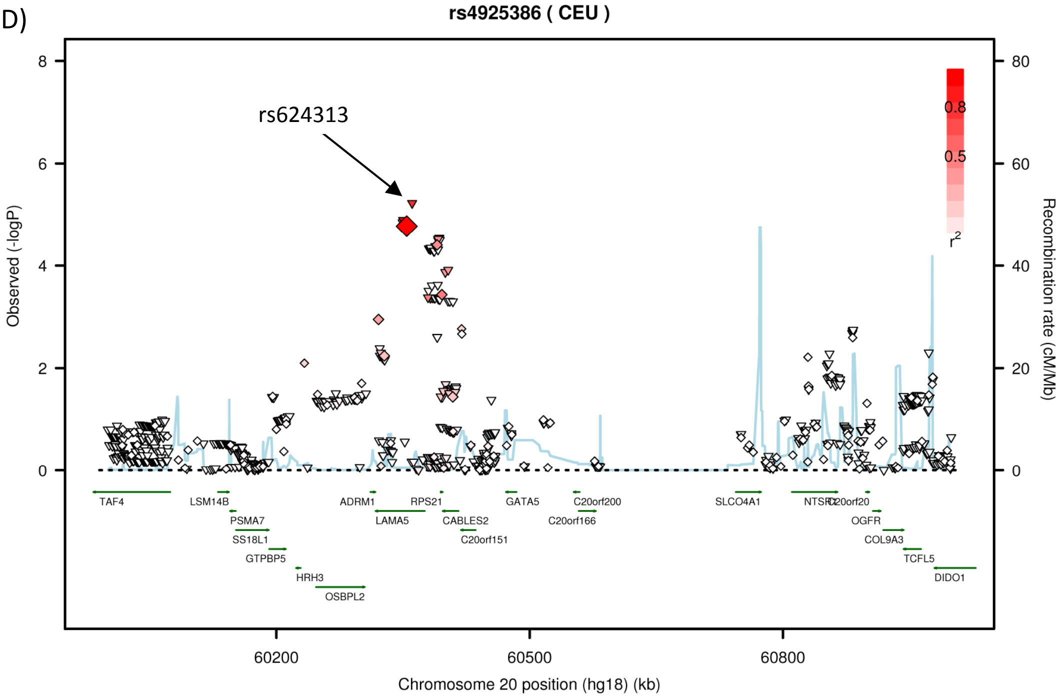
The imputed SNPs are shown as a triangle and genotyped SNPs as a diamond. The SNP named at the top of each plot is plotted as a large red diamond. LD relationships are only a guide as unfilled points are not necessarily independent of the named SNPs as this can indicate that they are not present in the HapMap panel used to determine LD. A) the region on chromosome 1, B) chromosome 3, C) chromosome 12 and D) chromosome 20. The plots were produced using the SNAP web based tool.



C)



D)



4.6 Fine-mapping the *CDH1* locus on chromosome 16

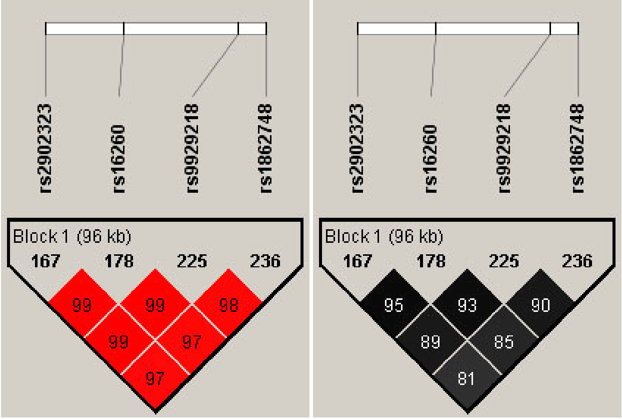
There were several promising candidate genes identified from the meta-analysis of the English and Scottish CRC GWA studies, discussed in Chapter 3, but I focussed on the

CDH1 region on chromosome 16. The *CDH1* gene codes for a calcium-dependent cell-cell adhesion molecule, which is thought to be a tumour suppressor gene. The expression of the protein is reportedly reduced in epithelial cancers (Takeichi 1991).

In a meta-analysis of EngP1, ScotP1, EngP2, Scot2 and VQ58, three SNPs showed suggestive association in this region under a fixed effects model. Two of these were within introns of the *CDH1* gene, rs9929218 (P=1.545x10^{-7}, OR=0.882) and rs1862748 (P=2.28x10^{-6}) and one, rs2902323, was in close proximity to *CDH3* (P=5.30x10^{-6}, OR=0.895). The SNPs rs9929218 and rs2902323 are highly correlated (r^2=0.89). All of these SNPs are in high LD with a functional SNP, rs16260, located in the promoter of *CDH1* (see Figure 4.19).

Figure 4.19 The LD between the most strongly associated chromosome 16 SNPs and rs16260

The pairwise LD between SNPs, measured by r^2 (left) and D' (right), which was calculated in Haploview using the EngP2 samples.

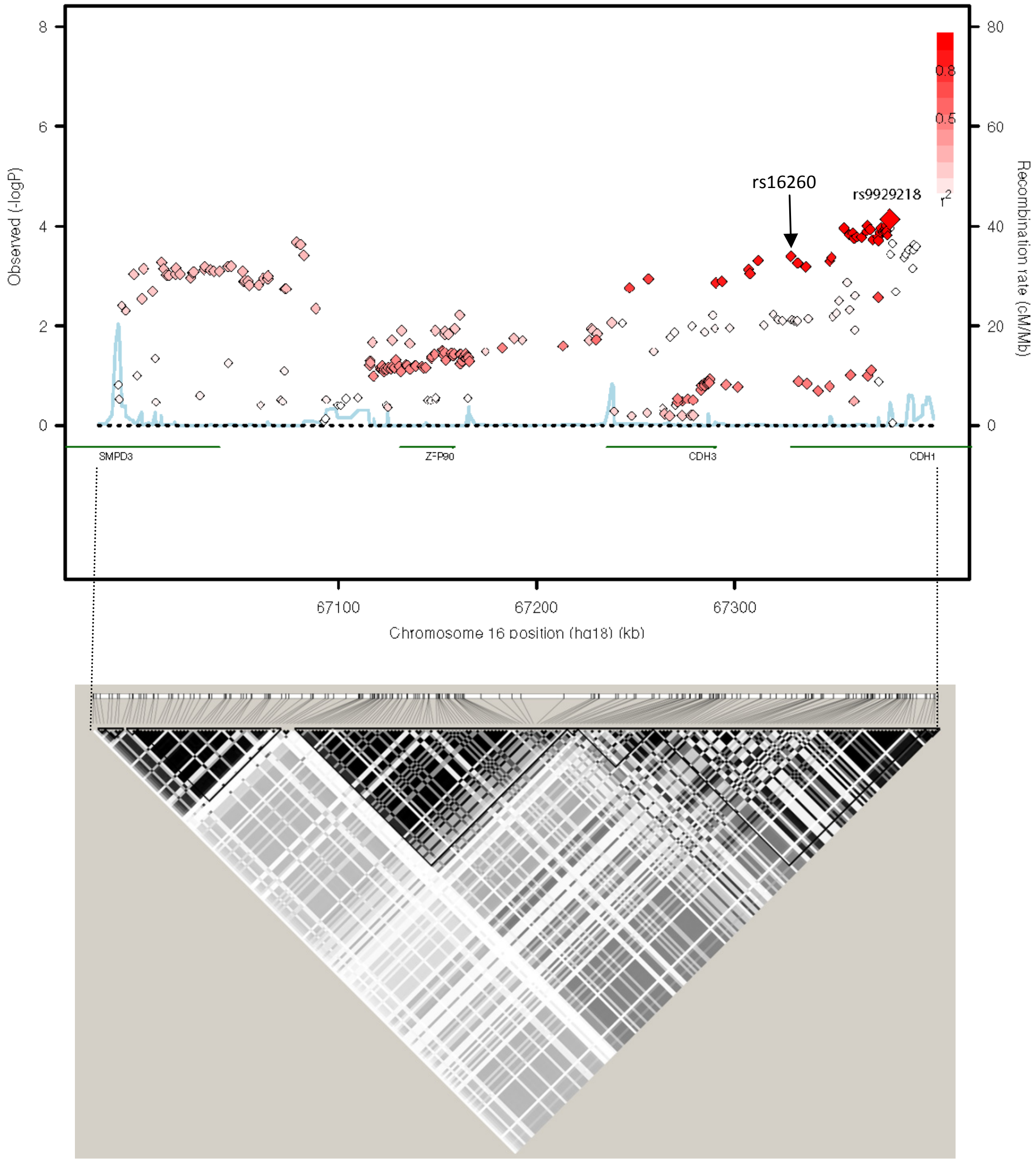


This SNP has previously been implicated in CRC (Porter *et al.* 2002) and prostate cancer (Verhage *et al.* 2002). It has been shown that the A (minor) allele of this SNP leads to decreased efficiency of *CDH1* transcription (Li *et al.* 2000). As rs16260 was not included on the Hap550 SNP array, I genotyped it in the EngP1 samples to check whether the association could be replicated and combined the results in a meta-analysis with that of the EngP2 and ScotP2 samples. The results for the three datasets upheld the correlation achieving P values of P=1.07x10^{-5} (OR=0.89) for rs9929218 and P=1.48x10^{-4} (OR=0.90) for rs16260. However, the CRC risk allele for rs16260 was the opposite allele to that associated with decreased CDH1 expression.

In order to determine if other potentially significant SNPs existed in the *CDH1* susceptibility locus and to attempt to fine map the region, we selected 243 SNPs between 66,988,860 and 67,391,657bp (genome build 36). These SNPs were genotyped in the large EngP2 and ScotP2 case control datasets. I combined the results in a meta analysis and the most significant SNPs were rs9929218 (P=7.25x10^{-5}, OR=0.88, 67,378,447bp), rs2961 (P=9.59x10^{-5}, OR=0.88, 67,376,404bp) and rs13339591 (P=9.70x10^{-5}, OR=0.88, 67,366,774bp). The P value of the SNP rs16260 was 4.01x10^{-4} (OR 0.89, 67,328,535bp, risk allele is C). For each of these SNPs, the ORs were calculated with reference to the minor allele and the risk allele is actually the major allele. The results for all 243 SNPs are plotted in the Figure 4.20, below, which illustrates that the most strongly associated SNPs are in strong pairwise LD with rs9929218, including rs16260 (see Figure 4.19). However, despite the higher P value in rs16260, it is still possible that this SNP is the causal allele. However, there may be multiple alleles in this locus that contribute to CRC risk.

Figure 4.20 The association results for the CDH1 region in EngP2 and ScotP2

This plot shows all 237 SNPs genotyped in the EngP2 and ScotP2 samples covering CDH1. The pair-wise r^2 LD values on this plot are all relative to the most significant SNP, rs9929218.



To determine whether the effects of the most strongly associated SNPs are indeed independent from rs16260, I performed a logistic regression analysis for the three most significant SNPs using the EngP2 and ScotP2 samples. The results, after correction for the effect of rs16260, suggest that none of the SNPs are independent of this SNP (rs9929218: $P_{logistic}$=0.01794, OR=0.735; rs2961: $P_{logistic}$=0.02939, OR=0.753 and rs13339591: $P_{logistic}$=0.03633, OR=0.756). Therefore, it is possible that the functional SNP, rs16260, is actually the causal allele.

4.6.1 Haplotype analysis of the *CDH1* region SNPs

I performed a haplotype analysis of the SNPs in this region, using PLINK (hap-logistic option), which considered haplotypes between two and ten SNPs with a frequency greater than 0.01. Individuals with more than 5% missing haplotypes were excluded. The analysis showed that carrying the risk alleles, CAAG, for each of these four SNPs increased CRC risk by 1.13 (P=1.78×10^{-4}, see Table 4.5).

Table 4.5 The frequencies for the low and high risk haplotypes of the chr16 SNPs

This analysis only considered the four most strongly associated SNPs. The alleles in the haplotype are given in the order of the SNPs in the first column.

SNPs	Haplotype	Freq (aff)	Freq (ctrl)	P	OR
rs16260, rs13339591, rs2961, rs9929218	CAAG	0.73	0.70	1.78×10^{-4}	1.13
	AGGA	0.26	0.28	3.07×10^{-4}	0.89

However, the most significant haplotype if all 237 typed SNPs are considered consisted of four SNPs, rs12929081, rs7186333, rs7186084, rs2059254, that are located between 67,372,629bp and 67,374,940bp (GACG, P=3.78×10^{-5}, OR=1.14, (Table 4.6). All of these SNPs are located in the intron region of CDH1, 46kb from rs16260.

Table 4.6 The frequencies for the most strongly associated haplotypes in the *CDH1* region

This analysis included all 237 typed SNPs of which the most strongly associated haplotype included just four SNPs. The alleles in the haplotype are given in the order of the SNPs in the first column.

SNPs	Haplotype	Freq (aff)	Freq (ctrl)	P	OR
rs12929081, rs7186333, rs7186084, rs2059254	GACG	0.7242	0.697	3.84×10^{-5}	1.14
	GTGA	0.2698	0.295	1.13×10^{-4}	0.88

4.6.2 Candidate gene screening of *CDH1*

All of the most strongly associated SNPs from the above analysis are located within introns and none of these SNPs would appear to have a direct functional effect on the expression or activity of *CDH1*, except rs16260 which is located in the promoter region of the gene. Therefore, I screened the exons of the *CDH1* gene for mutations that might explain the significant association of the tagging SNPs with disease. Ideally, a large number of samples are required to ensure that rare mutations are detected and to enable some investigation of the relative risk inferred on those that carry the mutation. In a balance between cost and power to detect a mutation in this gene, I screened 174 cases from the EngP1 dataset. All of these samples have a family history of disease, which should mean that any low penetrance gene mutations will be enriched in this group compared to those of unselected non-familial cases.

CDH1 has 16 exons, which I screened for variants using a combination of the LightScanner (Idaho Technology Inc.), which scans for sequence variations by measuring differences in the melting temperature of amplified DNA, and standard sequencing to identify sequence variation in any samples in which changes were found (see Chapter 2). If an exon contains more than one SNP, it is difficult to discern genetic differences between individuals using the LightScanner. Therefore, these exons were screened by sequencing. Both of these techniques require amplification of the sample DNA by PCR and the primers for these reactions are given in the Appendix. Any changes identified by the LightScanner were verified by sequencing of both the forward and reverse strands.

I detected 16 variants and to determine if these had been detected in control populations previously, I compared the physical positions with those of variants discovered as part of the 1000 genomes project (1KGP) using the March 2010 release available on the web based genome browser (at http://www.1000genomes.org). There are 500 population controls of European descent in 1KGP and the variants listed in the genome browser only cover those that have a frequency greater than 1%. None of novel variants listed below were detected in the 1KGP data and further analyses and sequencing of a much larger number of samples will be required to determine whether possessing these variants effects CRC risk. IKGP identified, with 500 samples, one non-synonymous variant, three synonymous coding variants and six intronic or 3' UTR (untranslated region) variants within or near the *CDH1* gene (67,328,696-67,426,943bp). I detected one non-synonymous variant, four synonymous coding variants and eleven intronic or 3' UTR variants (see Table 4.7). All of the novel variants detected in this screen were heterozygous. The only homozygous changes identified were known SNPs.

Table 4.7 The novel variants detected in the *CDH1* gene

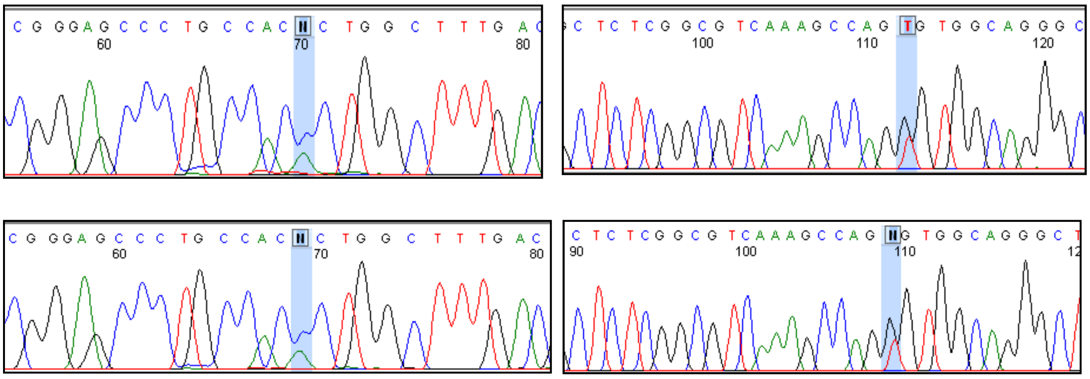
All 16 detected novel variants are given in the table, just one of which was non-synonymous and detected in two individuals. 'NS' indicates that the variant is non-synonymous and changes the amino acid and 'Syn' a synonymous variant.

ID	Site	Position (NCBI36)	Sequence	Type	Amino acid change	No of samples
6	In12	67,413,641	ACCTG[A/T]GTTTT	Intron	NA	1
21	Ex16	67,424,822	CAAAGA[C/T]CAGGAC	Syn	GAC(Asp) to GAT(Asp)	1
23	Ex16	67,424,804	GAACTC[C/T]TCAGAG	Syn	TCC(Ser) to TCT(Ser)	1
29	Ex16	67,424,720	TGACCC[C/T]ACAGCC	Syn	CCC(Pro) to CCT(Pro)	1
40	3'UTR Ex16	67,424,918	AGAGAG[G/T/A]CGGGCC	3'UTR	NA	1
41	3'UTR Ex16	67,424,951	ATGCAG[A/T]AATCAC	3'UTR	NA	1
42	3'UTR Ex16	67,424,934	GACCCA[T/A]GTGCTG	3'UTR	NA	1
51	In5	67,400,267	CTCTTA[G/A]AAGCTT	Intron	NA	2

55	Ex2	67,329,740	TGCCAC[C/A]CTGGCT	NS	CCT(Pro) to ACT(Thr)	2
56	Ex2	67,329,766	CTACAC[G/T]TTCACG	Syn	ACG(Thr) to ACT(Thr)	1
57	In2	67,329,838	GGTGTC[C/T]CTGGGC	Intron	NA	1
58	In2	67,329,833	CTGCCG[G/T]TGTCCC	Intron	NA	1
60	In2	67,328,879	AGAAAT[T/A]GCACTC	Intron	NA	1
62	In10	67,406,897	TTTTTAA[C/A]TTCATT	Intron	NA	1
63	In10	67,406,899	TTAACT[T/A]CATTGT	Intron	NA	2
67	In11	67,407,251	CATGGC[A/T]TTTTGT	Intron	NA	1

Figure 4.21 The forward and reverse sequences for the non-synonymous variant number 55

The sequences for the identified non-synonymous change in the two individuals that carried the variant are shown one above the other. The traces on the left show the forward strand and those on the right show the reverse.



The variant number 55 was the only non-synonymous coding variant discovered by this screen (Figure 4.21). It was detected in two individuals at a frequency of approximately 1% in the sample. I used the functional effect prediction tool, polyphen to predict whether this change was likely to affect the function of the protein (http://genetics.bwh.harvard.edu/pph/). However, the variant was predicted to be a benign change.

4.7 Discussion

In this chapter I have explored the use of imputation both to improve the overlap of SNPs between VQ58 and the other GWA datasets to allow meta-analysis and also to aid the fine-mapping of loci associated with CRC risk.

Attempts at identifying the causal variant underlying the association signals in this GWA study, and others, have proved difficult. In sequencing the exons of the *CDH1* gene on chromosome 16, I detected one non-synonymous coding variant in two samples from 174 samples studied, but the change was predicted to be benign. It is likely that sequencing the loci in a large numbers of individuals is required and that the signal is probably a combination of variants, which in some way affect expression of the protein. The results from this study showed that the functional SNP, rs16260, which is located in the promoter region of *CDH1* is associated with disease susceptibility, but the risk allele is the opposite allele to that found to decrease the expression of CDH1. It remains unclear where the causal variant lies.

The 174 samples chosen for screening were not selected based on genotype, and although the risk allele for the most strongly associated SNP in this locus is the major allele, it may still have been beneficial to enrich the sample for those individuals that carry the risk allele to increase the chance that a causal variant will be discovered.

It was clear from the imputation analyses for the VQ58 datasets that the reference panels chosen and the quality control criteria, such as removing SNPs with a low average maximum genotype probability, can affect the overall quality and the number of imputed SNPs available for subsequent analysis. The results of this work support the fact that quality of imputation can improve with an increased sample size in the reference panel, but also that the Hap300 genotypes are not ideal for imputation to HapMap and resulted in 408K fewer usable SNPs than the Hap550 SNP panel.

I have also shown the importance of imputing the same SNPs in both cases and controls within a datasets to avoid the introduction of bias into the results. The interpretation of imputed results should be treated with care and backed up with further genotyping to ensure that the finding is not an artefact. We were only able to successfully impute approximately 50% of the SNPs missing from the Hap300 array that were present on the Hap550, but even this allowed an additional 113,000 SNPs to be included in the meta-analysis. However, so far, all of the SNPs with a confirmed association with disease were actually genotyped in this dataset.

The imputation of untyped SNPs is unlikely to identify regions not found in the study data, but it is a valuable method for fine-mapping disease loci and identifying more strongly associated SNPs that may be able to refine the location of the causal variant. In hindsight, imputation of all available datasets to 1KGP P1 data, would have greatly increased the number of common variants included in the study and improved the fine-mapping of the regions to help locate causal variants. The 1KGP data provides five times the density of SNPs in HapMap phase II (Liu *et al.* 2010). This was especially evident in the plots shown for the chromosome 12 locus imputed to 1KGPP1/HapMap3 compared to the plot of the same region imputed to HapMap phase II. The June 2010 release of the 1KGP data is now available as a reference panel for IMPUTE. However, alternative methods for the detection of less common SNPs may still be required as this study is limited by the SNPs present on the Illumina chips used for genotyping as any imputed SNPs still need to be tagged by the study genotypes and less common variants with MAF<0.01 were not well tagged by the Hap550 (genome coverage is 87%). Further research is required to fully elucidate the missing heritability of CRC and determine the site and function of the causal variants that influence the risk of this disease.

Imputation increases the number of common SNPs included in the study and although we are unlikely to detect additional regions from those identified in the original GWA study, it increases the possibility of detecting the causal allele if it is common (and in HapMap) or detecting SNPs that are in higher LD with it. Multi-marker approaches have been shown using simulations by Spencer and colleagues who studied multi-marker tests including Imputation of untyped SNPs, which was found to confer a higher increase power than performing multi marker tests using only the genotyped SNPs on the arrays (Spencer *et al.* 2009). These types of studies show that calculations of power should be based on the chip being used and the LD of SNPs on that array by simulations rather than an analytical approach only taking into account allele frequency, and the number of SNPs and samples.

Chapter 5. Association analysis of the X chromosome

5.1 Introduction

So far I have only discussed CRC risk in terms of autosomal variation. This Chapter is dedicated to the analysis of variation on the X chromosome. There does not seem to be a gender bias in the cases of the datasets included in our study, although, there is a higher rate of CRC in the UK population in males (CRC rate of 70.2 in males and 56.6 in females per 100,000). It is important for completeness to determine whether common variants on the X chromosome influence CRC risk. Equally, the variants described so far in our GWA study explain approximately 8% of the excess familial risk and, together with the known high-penetrance Mendelian predisposition genes, do not fully explain the familial risk of CRC and it is plausible that risk variants also exist on the X chromosome.

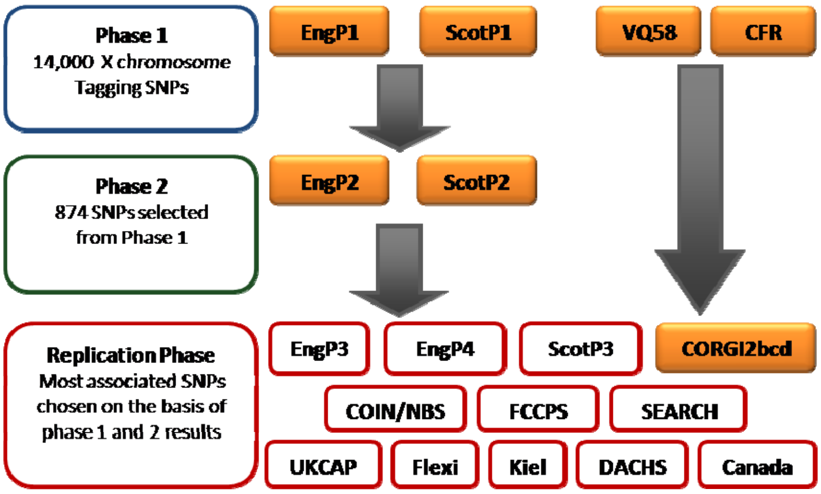
5.1.1 The X chromosome and cancer risk

Evidence relating the X chromosome to cancer risk includes studies in prostate cancer, which revealed an association to the variant rs5945572 at Xp11.22 and to a haplotype at Xp27.2 (Gudmundsson *et al.* 2008; Yaspan *et al.* 2008). The cancer gene census, originally published in 2004 (Futreal *et al.* 2004) contains 19 genes located on the X chromosome that have been linked to cancer. These include the transcription factor GATA protein binding 1 (*GATA1*), which has a role in erythroid development, the transcription factor forkhead fox O4 (*FOX04*), which is thought to play a role in development processes, SH2 domain containing 1A gene (*SH2D1A*), which encodes a protein involved in the bidirectional stimulation of B and T cells and the isoforms of septin-6, which is associated with acute myeloid leukaemia and belong to a family of

GTPases. Additionally, a novel tumour suppressor was identified, Wilms tumour suppressor (*WTX*, *FAM123B*), that down-regulates the WNT signalling pathway through the destruction of β-catenin and interacts with other proteins including APC and AXIN1 (Major *et al.* 2007). This pathway plays a major role in cancer progression as it regulates cell growth and differentiation.

5.2 Study design and sample datasets

Figure 5.1 The datasets included in the X chromosome analysis



The datasets included in this analysis are given in Figure 5.1. To assess the effect of X chromosome variants in the UK population in relation to CRC, I analysed the genotyping data for approximately 14,000 chrX tagSNPs from our four large case-control datasets, EngP1, ScotP1, CFR and VQ58 (CFR was genotyped for nearly 21,000 chrX SNPs). In order to increase the power of the study to detect variants by increasing the number of SNPs analysed, I implemented standard imputation methods using Impute v1 (described in the Materials and Methods Section 2.2.1.1) to predict all 64,622 SNPs in the HapMap phase II (r21 build 35) X chromosome reference panel to provide more SNPs for analysis and allow greater SNP overlap between datasets, which were genotyped on three different SNP arrays.

Imputed SNP genotypes were generated using IMPUTEv1 with the 'chrX' option The criteria for quality control of the imputed SNPs was the same as previously and consisted of a MAF greater than 0.01, information score greater than 0.5 (proper_info in SNPTEST), a maximum genotype probability greater than 0.9 for at least 95% of samples and HWE at P greater than 1x10^{-6}. The HWE was calculated using only the female datasets ('hwe' command in SNPTEST).

In addition to the GWA study datasets, 874 SNPs were genotyped on the X chromosome in EngP2 and ScotP2. Both of these datasets were genotyped for SNPs chosen primarily from the most strongly associated SNPs from the EngP1 and ScotP1 GWA studies, but an X chromosome specific analysis was not performed. Additional SNPs were included that were chosen based on identified candidate regions, copy number variations and non-synonymous SNPs that were not on the Hap550 arrays and only 481 of the 874 genotyped SNPs overlap with those genotyped in the other four datasets. As EngP2 and ScotP2 provide a further 5,670 and 4,063 samples, respectively, it was decided to add these datasets into the meta-analysis to increase the power to detect an association. I also attempted to impute the HapMap SNPs, to provide additional overlap. As expected, owing to the much lower density of genotyped SNPs, many SNPs did not impute successfully. However, a total of 3,719 and 3,736 SNPs passed quality control criteria in EngP2 and ScotP2, respectively.

Varying numbers of SNPs passed the QC process in each dataset and this number also differed between males and females within the same dataset (this is summarised for each dataset in Table 5.1). More SNPs were successfully imputed in the male samples. I assume that this is owing to the absence of heterozygote genotypes in these samples, which removes any difficulties in phasing the genotypes. I excluded SNPs that are not called in more than 5% of samples and as heterozygote genotypes are harder to call than homozygote genotypes, more female samples will not be called causing more SNPs to be excluded. Only SNPs that passed the quality control in both the males and females were included in the analysis. This led to 45,987 SNPs available for analysis.

Table 5.1 The number of SNPs after imputation in each dataset

The table shows the number of SNPs, with MAF<0.01, that passed the quality control criteria after imputation to the HapMap phase II X chromosome SNPs. The number of SNPs passing this threshold varies between males and females, but this is likely to be owing to the lack of heterozygotes in the males, which removes any phasing uncertainty leading to improved imputation performance. The intersection is the number of SNPs that overlap between males and females for each dataset. The SNP numbers include both genotyped and imputed SNPs.

Dataset	Total SNPs genotyped	Total number of SNPs after Imputation		
		Females	Males	Intersection
EngP1	12,186	48,939	53,561	48,243
ScotP1	12,428	49,220	54,051	48,698
CFR	20,282	53,586	56,608	52,957
Total Overlap	**11,199**			**45,987**
VQ58	8,296	40,132	47,723	40,022
Total Overlap	**7,934**			**38,163**
EngP2	794	3,721	4,443	3,719
ScotP2	794	3,738	4,469	3,736
Total Overlap	**481**			**2,920**

For the combined analyses, only SNPs genotyped or successfully imputed in the three datasets EngP1, ScotP1, and CFR were included. VQ58 was genotyped for fewer SNPs, compared EngP1, ScotP1 and CFR, and so the results from the meta-analysis include SNPs that are missing in VQ58. For the same reasons, this also applies to the EngP2 and ScotP2 datasets.

In order to attempt to replicate the effect of the most strongly associated SNPs, those that achieved a P value less than $1x10^{-4}$ were genotyped in the CORGI2bcd replication dataset, which consisted of 1092 controls and 588 cases. SNP genotyping was performed by my colleague Kimberley Howarth (primers are given in the Appendix).

5.3 Statistical Analysis of variation on the X chromosome

The X chromosome is slightly more complicated to analyse compared to the autosomes and therefore was not analysed as part of the original GWAS. As males have only one X chromosome and, therefore, do not have heterozygote genotypes, the data needs to be analysed differently. Males do not meet the normal assumption in GWA studies of HWE. There is a case for halving the allele count in males to reflect the absence of a second X chromosome, which leads to a reduction in power by effectively halving the sample size. However, the X chromosome is subject to random inactivation in females to prevent double dosage of expressed gene products. Therefore, in any cell only one allele will be expressed, which essentially allows us to treat males as homozygous females with regard to allele counting. The simplest way to deal with this, and the method I employed, was to analyse males and females separately to generate P values for association and then combine the results using a meta-analysis to produce single P values for each SNP (Jonathan Marchini, personal communication).

All analyses were conducted using the SNPTEST (v1) and META (v1) programs. The association statistics used in this analysis were the 'frequentist_additive_proper' P values, which are generated using the Cochran Armitage additive test. Males and females were analysed separately in SNPTEST using the 'exclude_samples' option to exclude the males or the females as appropriate. To obtain a single test result for each dataset, the male and female P values were combined using a fixed effects model in the program META (which incorporates standard error, sample size and effect size beta values into the analysis). However, the meta-analysis of all datasets was performed on the individual male and female analysis results, for example, the EngP1, ScotP1 and CFR analysis was a combined analysis of six groups.

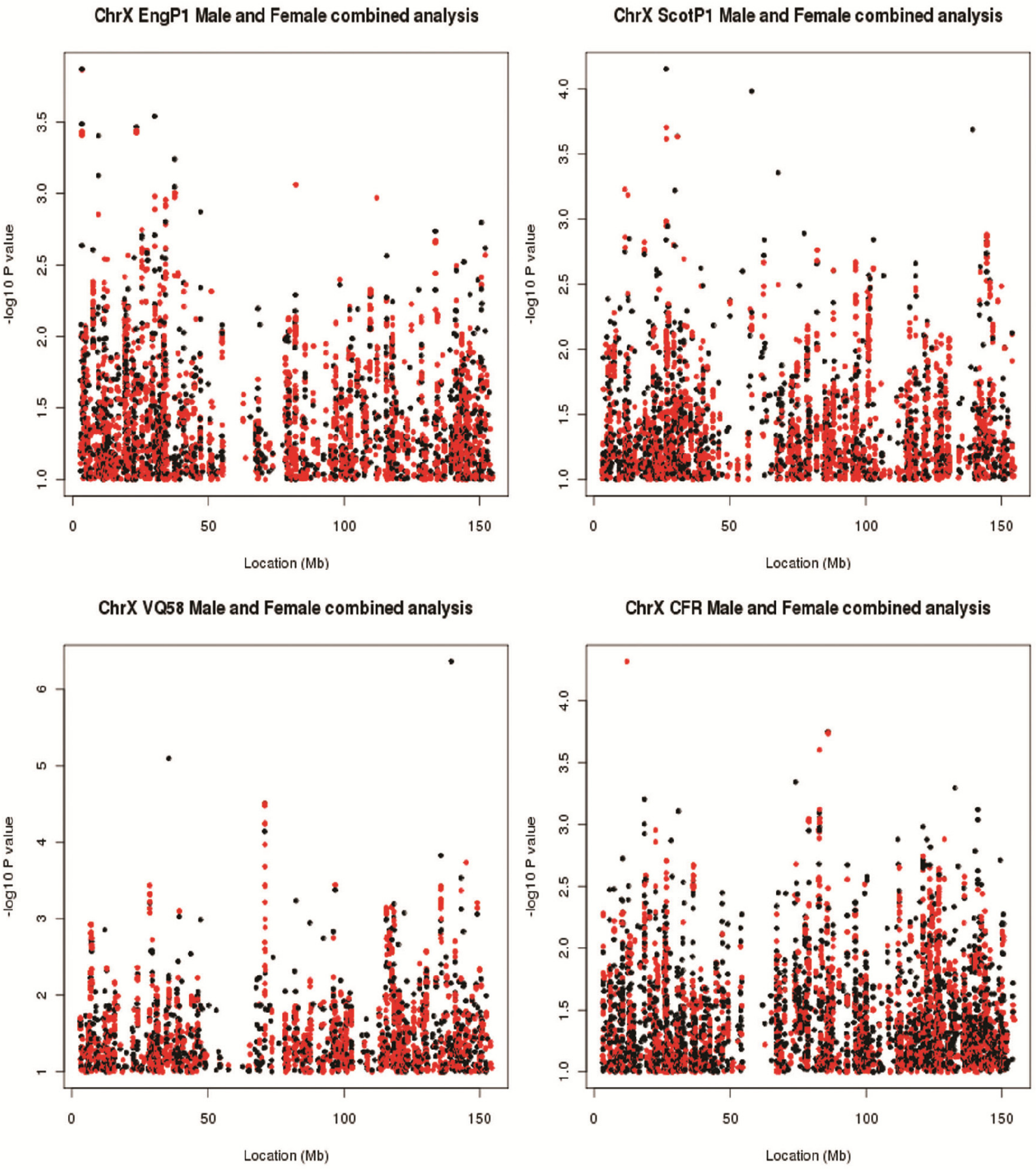
5.4 X Chromosome Results

5.4.1 Genotyped and imputed SNPs across the X chromosome

Each dataset was initially analysed independently to allow quality control measures to be applied separately. The male and female results for each dataset were then combined in a meta-analysis. The significance level taking into account multiple testing of only the X chromosome 45,987 SNPs would be $1.09\text{x}10^{-6}$ using a Bonferroni correction. However in reality a genome-wide significance level is probably more appropriate. In considering a meta-analysis of the males and female results in each dataset, the only SNP to reach genome-wide significance was the genotyped SNP rs4824847 at 139,474,356bp, which achieved a P value of $4.31\text{x}10^{-7}$ in VQ58 (beta=0.579 and se=0.115, see Figure 5.2). However, this SNP was not replicated in the other datasets. No SNPs reached significance in any dataset when the males or the females were analysed independently.

Figure 5.2 A summary of individual association analysis results for each dataset

The plots below show the $-\log_{10}$ P values plotted against SNP location for imputed (red) and genotyped (black) SNPs on the X chromosome for the four main GWA study datasets, EngP1, ScotP1, CFR and VQ58. The plots show that few of the most strongly associated SNPs overlap between datasets. The lowest P values were found in the VQ58 dataset (rs4824847), but this was not replicated in the other datasets.



5.4.2 Meta-analysis of the four datasets

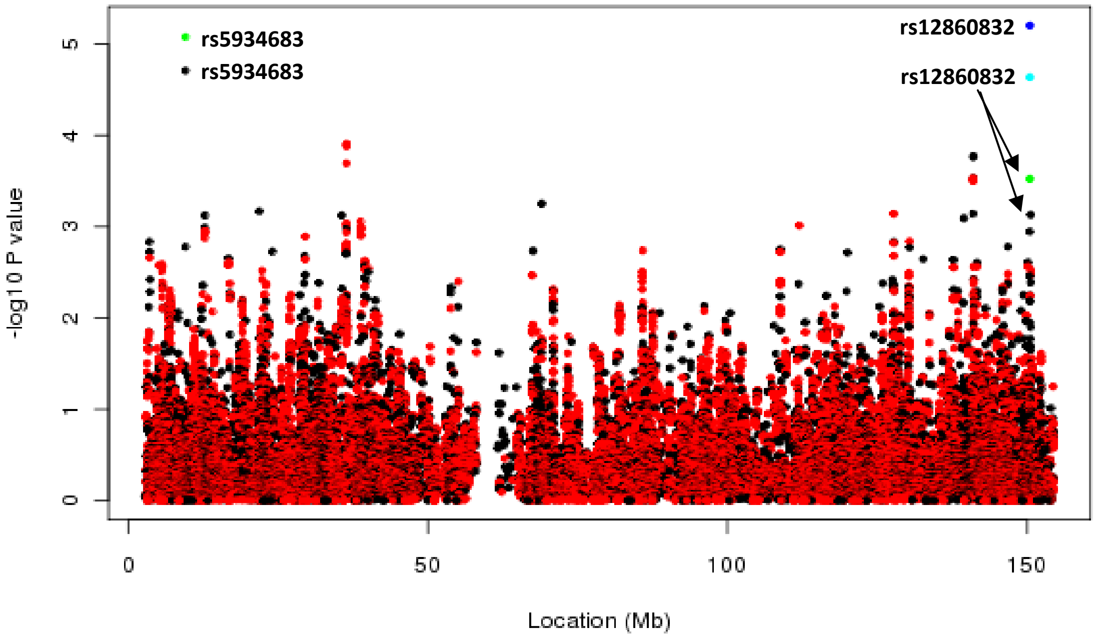
The meta-analysis of the chromosome X SNPs combining the EngP1, EngP2, ScotP1, ScotP2 and VQ58 datasets, identified two SNPs showing evidence of association. The results from this analysis are shown in Figure 5.3 and also include imputed SNPs. The most strongly associated SNP from this analysis was rs5934683 at 9,561,210bp (P=1.95x10^{-5}, beta=0.07, SE=0.017). This SNP was genotyped in the CORGI2bcd replication phase and included in a meta-analysis to achieve an overall P value of 8.38x10^{-6} (beta=0.075, SE=0.017, see Table 5.2 and Figure 5.6).

The second associated SNP identified by this analysis was rs12860832 at 150,482,104bp, which was originally identified in an analysis of just the genotyped SNPs (P=2.31x10^{-5}, beta=-0.146, se=0.034, see cyan coloured point in Figure 5.3). This SNP was not genotyped in EngP2, ScotP2 or VQ58. It was imputed in VQ58, however, when these results were included in the meta-analysis the signal was not improved (P=6.79x10^{-4}, beta=-0.09, SE=0.027).

Owing to the low P values achieved without VQ58, we genotyped the SNP in this dataset to ensure that the result was not influenced by the imputation. The genotyped results were used in the combined analysis. This SNP was then genotyped in the CORGI2bcd replication dataset and the results of the overall meta-analysis using genotyped data in VQ58 gave a P value of 3.00x10^{-4} (beta=-0.089, se=0.025).

Figure 5.3 The combined analysis for the X chromosome for EngP1, EngP2, ScotP1, ScotP2, CFR, VQ58g

The results of the meta-analysis of the all available datasets are illustrated and show the results for SNPs that are present in EngP1, ScotP1 and CFR. For SNPs present in EngP2, ScotP2 and VQ58, these groups are included in the analysis. The most strongly associated SNP was rs5934683, which was genotyped in all 6 datasets. The second SNP taken forward for replication was rs12860832, which was originally detected in an analysis of the genotyped SNPs. The SNP is not genotyped in VQ58 or EngP2/ScotP2 and the original P value and the replication are shown as a cyan and blue point, respectively. The black and green points shown by the arrows relate to the analysis with VQ58 included, see text. Black points show genotyped SNPs, red points show imputed SNPs and the green points show the meta-analysis result for the SNPs genotyped in CORGI2bcd for replication.



The results for both of the identified SNPs separated by gender and dataset are summarised in Figures 5.4 and 5.5. Across five case/control datasets these data, despite the VQ58 males in rs12860832, provide good evidence of an association of two SNPs on the X chromosome with CRC risk.

Figure 5.4 The results for rs5934683 separated by group

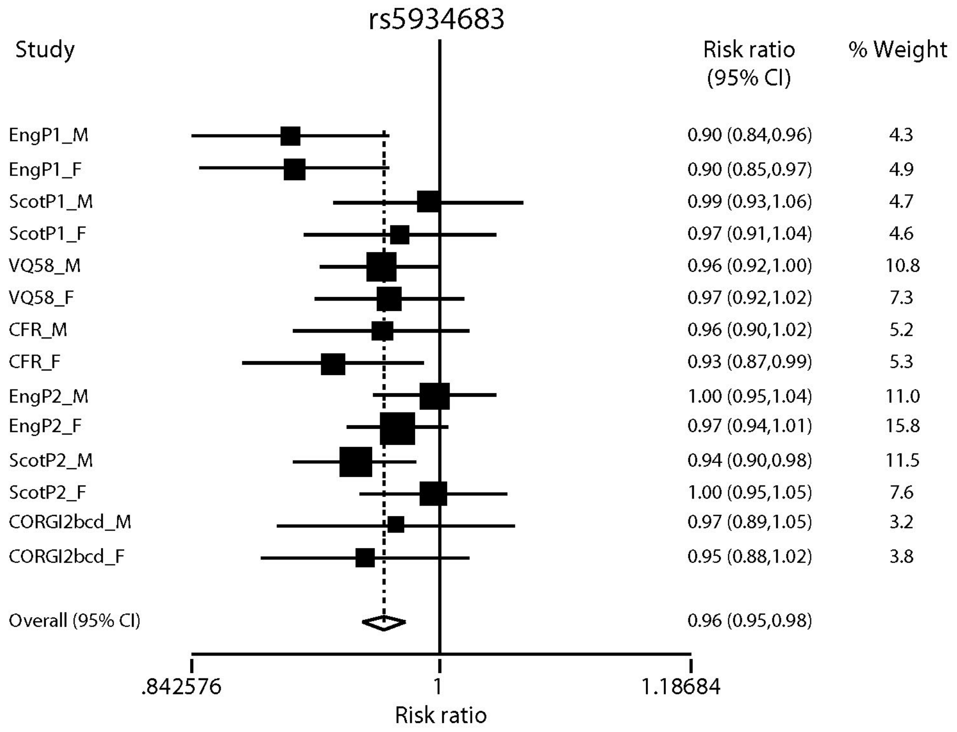


Figure 5.5 The results for rs12860832 separated by group

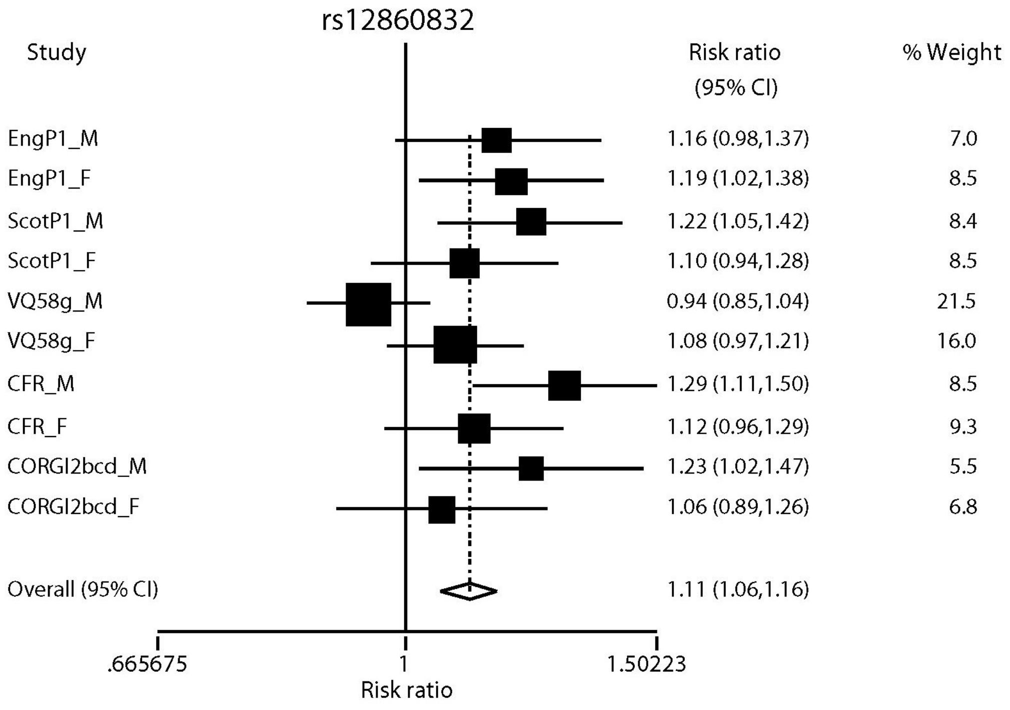


Table 5.2 Summary statistics for the two most strongly associated SNPs

The summary statistics for the two identified SNPs in each individual dataset. The combined MAF is the mean of the male and female allele frequencies. The P values are the frequentist additive test in SNPTEST (Cochran-Armitage test) and the model parameter, betas, estimates and their standard errors are also given. The beta values are calculated with reference to the B allele (where the allele A is coded as 0 and allele B is coded as 1, the beta is an estimate of the increase in log-odds that can be attributed to each copy of the B allele). The minor allele is the A allele for both SNPs. The HWE P value is calculated from the controls of the female group for each dataset. VQ58g indicates genotyped data that contains additional cases that were not included in VQ58i. The P value for between-study heterogeneity was greater than 0.05 for each of these SNPs.

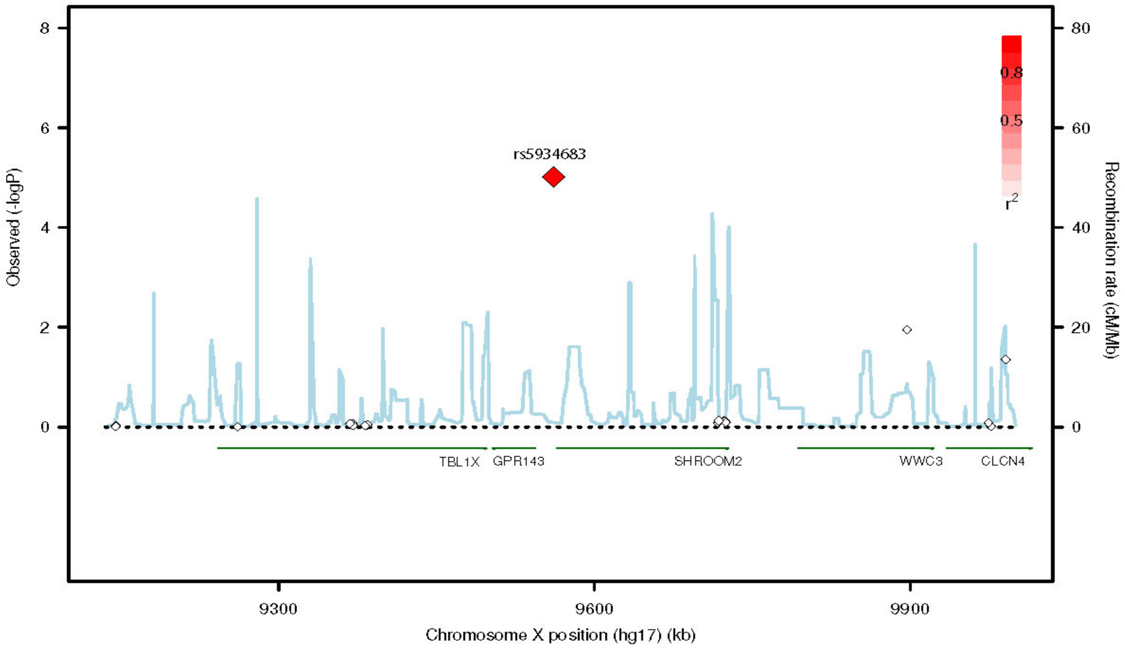
SNP	Position (Mb)	Allele (A/B)	Group	Gender	Genotype Counts						P	Beta	SE	HWE P (ctrls)	OR	95% CI	MAF Cases	MAF Controls
					Cases			Controls										
					AA	AB	BB	AA	AB	BB								
rs5934683	9,561,210	C/T	EngP1	M	272	0	160	294	0	127	0.033	0.154	0.072	-	1.36	1.11-1.67	0.370	0.302
				F	181	240	65	240	214	53	2.23x10^{-3}	0.289	0.095	0.61	1.33	1.11-1.60	0.381	0.316
				Combined							3.93x10^{-4}	0.204	0.058				0.376	0.309
			ScotP1	M	317	0	178	332	0	182	0.855	0.01	0.07	-	1.02	0.85-1.23	0.360	0.354
				F	195	226	58	212	220	55	0.397	0.08	0.10	0.92	1.08	0.90-1.31	0.357	0.339
				Combined							0.530	0.034	0.054				0.358	0.346
			VQ58	M	592	0	299	963	0	428	0.157	0.07	0.05	-	1.14	1.00-1.29	0.336	0.308
				F	223	245	64	581	586	132	0.169	0.11	0.08	0.41	1.11	0.96-1.29	0.351	0.327
				Combined							0.055	0.076	0.040				0.343	0.317
			CFR	M	396	0	220	320	0	158	0.354	0.06	0.06	-	1.13	0.94-1.34	0.357	0.331
				F	222	272	80	227	243	49	0.018	0.22	0.09	0.20	1.23	1.03-1.47	0.376	0.329
				Combined							0.034	0.111	0.052				0.367	0.330
			EngP2	M	782	0	429	742	0	402	0.880	0.01	0.04	-	1.01	0.90-1.14	0.354	0.351
				F	673	759	185	759	694	197	0.100635	0.085	0.052	0.05	1.09	0.98-1.21	0.349	0.330
				Combined							0.245	0.039	0.033				0.352	0.341
			ScotP2	M	751	0	457	797	0	412	0.055	0.08	0.04	-	1.18	1.05-1.32	0.378	0.341

				F	340	348	104	356	372	103	0.849	0.01	0.07	0.70	1.01	0.88-1.17	0.351	0.348
				Combined							0.079	0.064	0.037				0.365	0.344
			CORGI2bcd	M	169	0	105	306	0	175	0.596	0.04	0.08	-	1.09	0.87-1.35	0.383	0.364
				F	128	131	38	251	231	52	0.149	0.16	0.11	1.00	1.17	0.95-1.45	0.348	0.314
				Combined							0.202	0.08	0.06					
Overall meta-analysis of all datasets											**$8.38\text{x}10^{-6}$**	**0.075**	**0.017**					
rs12860832	150,482,104	A/G	EngP1	M	113	0	319	95	0	326	0.222	-0.10	0.08	-	0.82	0.66-1.03	0.262	0.226
				F	33	204	250	32	173	302	0.026	-0.23	0.10	0.32	0.80	0.65-0.97	0.277	0.234
				Combined							0.020	-0.147	0.063				0.269	0.230
			ScotP1	M	138	0	357	117	0	397	0.062	-0.14	0.07	-	0.76	0.62-0.93	0.279	0.228
				F	39	176	264	33	169	285	0.240	-0.12	0.10	0.26	0.88	0.72-1.08	0.265	0.241
				Combined							0.028	-0.130	0.059				0.272	0.234
			VQ58g	M	234	0	722	362	0	1029	0.397	0.04	0.05	-	1.09	0.95-1.24	0.245	0.260
				F	54	247	339	88	489	721	0.175	-0.10	0.08	0.66	0.90	0.77-1.04	0.277	0.256
				Combined							0.363	-0.120	0.132				0.261	0.258
			CFR	M	174	0	435	104	0	367	0.016	-0.17	0.07	-	0.71	0.58-0.86	0.286	0.221
				F	39	222	303	32	181	301	0.140	-0.15	0.10	0.47	0.86	0.71-1.05	0.266	0.238
				Combined							$4.84\text{x}10^{-3}$	-0.159	0.057				0.276	0.230
			CORGI2bcd	M	73	0	200	105	0	377	0.123	-0.14	0.09	-	0.76	0.60-0.97	0.267	0.218
				F	22	109	172	25	211	311	0.518	-0.08	0.12	0.19	0.93	0.74-1.17	0.252	0.239
				Combined							0.105	-0.12	0.07					
Overall meta-analysis of all datasets											**$3.00\text{x}10^{-4}$**	**-0.089**	**0.025**					

The detected SNPs are both in regions with high levels of recombination. The SNP rs5934683 (Figure 5.6) is located between the genes *GPR143* (G protein coupled receptor 143), which is involved in intracellular signal transduction, and *SHROOM2* (homolog of Xenopus apical protein), which is implicated in amiloride sensitive sodium channel activation. Both of these genes are strong candidates for the visual disorder ocular albinism type 1.

Figure 5.6 The position and r^2 values for rs5934683 and surrounding SNPs

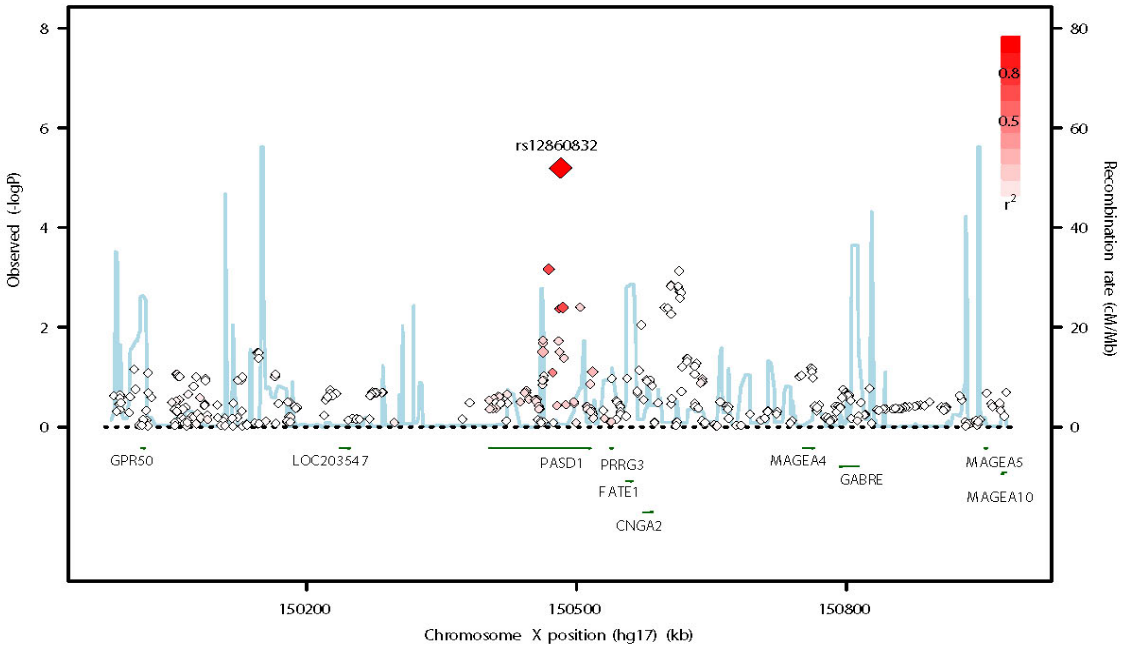
The plot shows the region surrounding the identified SNP, including the recombination rate plotted in blue and nearby genes. None of the other SNPs included in this analysis are in LD with rs5934683, but it is flanked by loci with high recombination rates. The r^2 was calculated using the HapMap CEU data. The SNP in located between the genes *SHROOM2* and *GPR143*.



The SNP rs12860832 is located within an intron of the gene *PASD1* (PAS domain containing 1), which is as a transcription factor (Figure 5.7). The PASD1 protein was identified as a potential immunotherapeutic agent as it is expressed in diffuse large B cell lymphoma (DLBCL) and other haematological cancers, but absent from normal tissue (Cooper *et al.* 2006; Sahota *et al.* 2006). PASD1 is a cancer-associated antigen that has been reported to stimulate cytotoxic T-cell response in tumours (Ait-Tahar *et al.* 2009).

Figure 5.7 The position and r^2 values for rs12860832 and surrounding SNPs

The plot shows the region surrounding the identified SNP, including the recombination rate plotted in blue and nearby genes. This SNP is also located between two loci with high recombination rates and there are few additional SNPs in high LD with the identified SNP. The SNP is located within an intron of the *PASD1* gene. The r^2 was calculated using the HapMap CEU data.



5.4.3 rs12860832 and issues with VQ58

The results for the rs12860832 showed that for all datasets the effect is in the same direction, except in the VQ58 males. This appeared to be due to the effect of the SNP being in the opposite direction in VQ58 males compared to the other groups.

Although the MAF for this SNP in the VQ58 cases is similar to the other datasets, the MAF in the controls is higher at 0.267 compared to the MAF of 0.230 seen in the other control groups.

As this SNP had been both imputed and genotyped in this dataset, I compared the results to determine if there was a difference that could be a result of genotyping error. However, both sets of results gave similar association results for the males (P=0.380, beta=0.04, se=0.05 when imputed and P=0.397, beta=0.04, se=0.05 when genotyped). The MAF in the typed males was 0.245 in cases and 0.260 in controls, which was comparable to that in the imputed data (0.25 in the cases and 0.27 in the controls). To determine the concordance of the imputed and genotyped data, I compared the calls in the VQ58 cases using only genotypes with a maximum genotype probability greater than 0.9 and only those samples with a called genotype (n=1203, see Table 5.3).

Table 5.3 Concordance between imputed and genotyped data in VQ58 cases for rs12860832

This data compares the genotype counts for rs12860832 when directly genotyped and when imputed in the VQ58 cases. Not all samples could be genotyped and the counts only include samples where DNA was available (1,203). As the imputation process imputes SNPs that are missing from the data owing to failed genotyping and as some genotyped SNPs will fail to impute, the concordance between imputed and genotyped datasets over all samples will not always provide a good representation of imputation quality. Therefore, I removed any sample that failed in either the imputed or genotyped datasets and calculated the concordance of the called genotypes, which showed a good level of concordance (95%). Called genotypes in the imputed dataset are based on a threshold maximum genotype probability of 0.9.

Genotyped/ Imputed	AA	AG	GG	Fails	Concordance
Genotyped	215	174	814	-	0.946
Imputed	227	191	785	-	

5.5 Discussion

This study has identified two SNPs on the X chromosome showing good evidence of an association with CRC risk, despite the loss in power caused by separating the males and females into separate groups and meta-analysing the results.

The results of the individual group analyses show that the effect of each of these SNPs is very small with most of the groups having ORs with 95% confidence intervals that span one. The SNP rs5934683 shows the largest effect in the family history enriched EngP1 dataset, which may be a factor influencing this association. Although the effect of each of the SNPs was replicated when genotyped in additional samples, neither reached genome wide significance and both require further genotyping in additional samples, such as the full replication dataset that has been utilised for the SNPs previously identified in this GWA study. The SNP rs12860832 achieved a P value of $5.53x10^{-4}$ without the addition of the EngP2 or ScotP2 datasets and needs to be genotyped in these samples to help confirm this association.

In contrast to the imputation results in previous chapters, the imputation of HapMap phase II SNPs on the X chromosome did not identify any SNPs more strongly associated with disease than those already genotyped. However, the study did demonstrate that imputation of rs12860832 in the VQ58 dataset to allow inclusion of this dataset into the analysis produced comparable association results using imputed or genotyped data (95% concordance). There remains a question mark over the difference in allele frequency for this SNP in the VQ58 controls compared to the other datasets included in this study. It is difficult to say for certain, but the imputed and genotyped data showed a comparable allele frequency indicating that this may be a chance difference between samples rather than a genotyping error.

Both SNPs indentified in this study are promising candidates for CRC risk, even though they are not located in regions with genes obviously linked to CRC development. However, both need to be genotyped in additional samples, such as the full replication phases, to determine whether these SNPs reach genome-wide significance to confirm the association.

Chapter 6. The analysis of runs of homozygosity and CRC risk

6.1 Introduction

The aim of this study was to use the high density genotype data generated in the GWA study to evaluate the presence of runs of homozygosity (ROHs) in an unrelated outbred case control dataset and determine whether increased homozygosity is associated with disease status. Although numerous susceptibility SNPs have been identified through GWA studies, no recessively acting SNPs have yet been reported, which is perhaps an indication of a lack of power to detect them in this type of study. However, the effects of these susceptibility alleles added to the known high-penetrance Mendelian mutations are a long way from explaining the total variance in risk. The remainder may be explained by other variations, such as SNPs not tagged in the arrays, rare variants, copy number variations, recessive alleles, or haplotypes that are not detected by single SNP analyses.

6.1.1 Homozygosity and inbreeding

Homozygosity mapping has been an important method for the detection of causal mutations in rare recessive diseases using inbred families where affected individuals are very likely to share the same mutation that is inherited through shared segments that are identical by descent (IBD). Between 1995 and 2003, 200 studies were published that described homozygosity mapping as a method of detecting disease genes (Botstein and Risch 2003).

The effect of long regions of homozygosity, caused by inbreeding, has been well studied owing to the detrimental effects on health that were observed in inbred families. In 1955, Penrose *et al.* first wrote about the advantages of heterozygosity seen in animal species and their applicability to humans in the context of various complex traits (Penrose 1955). Since then, a number of studies have used inbreeding to provide evidence for a recessive basis of cancer. This evidence has largely come from the analysis of large families with a high degree of inbreeding. Lebel and Gallagher studied a large 1000 member family with a high number of CRC cases (Lebel and Gallagher 1989). CRC affected 14 members of this family and 13 of these were the offspring of closely related parents. The observation of increased cancer risk has also been noted in some isolated populations such as those from 14 neighbouring villages in the Middle Dalmatian islands in Croatia (Rudan *et al.* 2003). It was estimated in 2001 that on average each person carries 500-1200 low frequency deleterious recessive alleles as heterozygotes (Fay *et al.* 2001). However, the authors suggest that in the presence of inbreeding many of these may become homozygous leading to large effects in individuals and an increase in the risk of complex diseases including cancer.

Incidentally, other studies have reported that inbreeding can actually decrease cancer risk. For example, a case control study from the United Arab Emirates reported a significant correlation (P<0.001) between increased inbreeding measured by the inbreeding coefficient and reduced cancer risk in a sample consisting of 391 cases and 378 matched controls (Denic *et al.* 2007). Although the sample size is small and it is possible that by chance the authors have selected a population with a low average cancer risk. Inbreeding should only increase cancer risk if it results in an increase in homozygous risk variants, if the frequency of these variants is very low and the variants are relatively recent then inbreeding is unlikely to increase cancer risk. The results suggest that the effects of inbreeding may vary between populations or be dependent on specific diseases.

6.1.2 Runs of Homozygosity in outbred individuals

Runs of homozygosity are not solely caused by inbreeding events. There are a number of alternative explanations for increased homozygosity including linkage disequilibrium (including long range LD), where correlated SNPs cause an increase in short homozygous segments, heterozygous deletion and chromosomal abnormalities such as inversions and uniparental disomy (UPD), which is where an individual receives both copies of a chromosome, or chromosome region, from just one parent, or hemizygous deletion caused by copy number variation (Wang *et al.* 2009).

Broman and Weber had identified individuals with long regions of continuous homozygous markers during the construction of genetic maps using individuals from CEPH (Broman and Weber 1999). The markers were short tandem-repeat polymorphisms, which are much more informative, but much less dense than SNPs, but had been genotyped in eight of the CEPH families. The authors suggested that the autozygous regions were possibly due to increased levels of inbreeding in past generations and that this could have utility in mapping regions or haplotypes that are shared among affected individuals, but absent from controls. Autozygosity occurs when both parents provide to their offspring a stretch of DNA that is identical by descent (IBD). The length of this region can give an indication of the degree of relatedness between the parents (Wang *et al.* 2009). Therefore, in outbred populations, individuals should be separated from a common ancestor by many generations and so shared regions should be split up by recombination events leading to the presence of only very short autozygous regions (Gibson *et al.* 2006).

In a study on ROHs in European populations, using genome-wide SNP data from 2,618 individuals using 289,738 SNPs on the Illumina HumanHap300 array, it was demonstrated that ROHs longer the 4Mb are often observed in outbred individuals (McQuillan *et al.* 2008). This study also showed that ROHs greater than 1.5Mb in length could be used to effectively distinguish between different populations. For this reason, any population stratification within a dataset that is used for studying the difference in ROHs between cases and controls could dramatically skew the results.

In a relatively recent study of haplotype structure, using 1411 samples, Curtis and colleagues demonstrated that stretches of homozygous SNPs spanning more than 1Mb were actually quite common across the genome (Curtis *et al.* 2008). The authors reported detecting these runs of homozygosity, covering an average of 73 SNPs, in 36% of samples. Owing to the seemingly non-random distribution of the ROHs, it was surmised that they could be caused by haplotypes that are common in the population leading to both parents giving the same haplotype to their offspring. The identification of haplotypes that are more frequent in cases than controls could be used to discover regions associated with disease.

6.1.3 Levels of homozygosity and CRC

Recently, two studies have reported evidence which seems to support the hypothesis that increased germline homozygosity is associated with an increased risk of cancer. The first was an investigation into loss of heterozygosity and allelic imbalance using 345 microsatellite markers, that were equally spaced across the genome, in paired tumour/normal DNA samples (Assie *et al.* 2008). This study compared 385 patients of north/west European ancestry who were diagnosed with cancer (147 with breast, 116 with prostate and 122 with head and neck cancer) with ethnically matched controls from the Cooperative Human Linkage Centre study (Murray *et al.* 1994). The authors measured levels of homozygosity by calculating the frequency of homozygosity at microsatellite markers in cases and controls and then measured LOH at sites of high homozygosity frequency in the tumours of the cases. The results showed a significant increase in the frequency of homozygosity of cases compared to controls at 114 loci. However, this study looked at the frequency of homozygosity at each individual marker to identify loci that were homozygous more often than would be expected by chance and not runs of homozygous SNPs. This method is a crude approximation of overall homozygosity as although microsatellites are more polymorphic than SNPs, they are far less dense across the genome.

In the second study the authors analysed levels of homozygosity based on the Affymetrix 50K *Xba*1 SNP array and focussed on CRC (Bacolod *et al.* 2008). The study population consisted of 74 CRC patients, with an average age of 66 years, whose germline DNA was extracted from the non-cancerous tissues of snap-frozen tumour samples. The controls were formed from two groups, 146 samples from the age-related macular degeneration (AMD) study and 118 samples from National heart, lung and blood institute Framingham Heart study.

The authors analysed runs of consecutive homozygous SNPs that covered more than 4Mb and included more than 50 homozygous SNPs and reported that cases had significantly more homozygous segments than controls ($P=1.28x10^{-5}$ compared with AMD controls and $P=1.13x10^{-5}$ when compared with Framingham controls). At least one homozygous segment that met the above criteria was detected in 62% of cases, and 29% and 36% of the Framingham and AMD controls, respectively. Additionally, the total length of detected segments in each sample was on average longer in cases compared to controls. However, a number of cases that were of Ashkenazi Jewish descent were included in the analysis and may have skewed the results owing to mismatched cases and controls. This population is known to have somewhat higher levels of inbreeding and therefore more homozygous regions than non-Jewish samples. The results from this study led the group to propose a model for the importance of homozygosity in cancer progression (Bacolod *et al.* 2009).

Together the studies detailed above support the hypothesis that several loci with low-penetrance, recessively-acting alleles that are not detected by the current GWA study approaches could contribute to cancer susceptibility in outbred populations. It could be that the frequency of these alleles is too low to be detected, that alleles are heterogeneous in the population or that they are simply not in LD with the SNPs on the currently available arrays. It was appealing to investigate ROHs in our reasonably large EngP1 case/control dataset, for which pedigree structure was known for the cases and individuals had been genotyped on the Illumina Hap550 genome-wide SNP array. As no definitive method existed for the analysis of such data, I have used several different methods to compare homozygosity in cases and controls.

6.2 Study design

6.2.1 Study samples

For this study, I analysed the EngP1 case control dataset, which consisted of 921 cases with confirmed CRC or advanced colorectal neoplasia (433 males and 488 females) and 929 healthy controls (422 males and 507 females). The VQ58 dataset was used as a replication dataset for aspects of this study. The 1958 birth cohort included in VQ58 was the WTCCC1 dataset genotyped on the Illumina HumanHap550 SNP array, which consisted of 1,438 samples.

6.2.2 SNP panels and quality control

The data used in this study is the same as that used for the GWA study described above. Only autosomal SNPs were included these analyses. Additionally, SNPs were excluded if the minor allele frequency was less than 5% in our samples and there was any deviation from HWE in either cases or controls ($P<1x10^{-5}$). The remaining 486,303 SNPs, which I refer to as the 500K panel, were used for the detection of runs of homozygosity in this study.

As long stretches of homozygous SNPs can be caused by LD between markers, it was prudent to remove SNPs in high pairwise LD and perform the analysis on a panel of independent SNPs to improve the detection of truly autozygous regions. Therefore, in order to exclude ROHs caused by LD between SNPs, I also performed the analyses on a ‘low-LD’ panel of 30,307 SNPs. The SNPs were chosen by pruning the 500K panel based on pairwise LD using the PLINK SNP pruning function (indep-pairwise) (Purcell *et al.* 2007), with a window size of 50 SNPs, a step size of 5 SNPs and pairwise r^2 threshold of 0.1, to produce a list of essentially independent SNPs.

6.2.3 Detection of ROHs

Runs of homozygosity were detected using tools available in PLINK v1.05 (Purcell *et al.* 2007). The PLINK ROH tool is accessed using the ‘homozyg’ command, which detects long stretches of homozygous SNPs from whole genome SNP genotype data. The function works simply by moving a sliding window, of a predefined number of SNPs (50 by default), across the genome and then at each window position determines whether the required level of homozygosity is reached. For each SNP, the proportion of homozygous windows that overlap the position is calculated and then used to call segments as ROHs if the minimum criteria are met.

The default function parameters and minimum ROH criteria are set appropriately for dense panels of genome wide SNPs, but were altered as described below. The number of heterozygotes permitted within a window was set to 2% (1 per 50 SNP window) and the number of permitted missing calls was set to 5 within a window. These settings were chosen to attempt to prevent an underestimation of ROH number and size caused by runs of truly homozygous SNPs being broken by the presence of a miss-called heterozygote or an uncalled genotype. Heterozygous SNPs could also be caused by mutation or gene conversion (Broman and Weber 1999).

I have used various criteria for calling segments as an ROH in order to analyse the data based on size defined either by the number of SNPs or the number of kb that the segment covered. To this end, I repeated the analysis with several different values for the 'homozyg-snp' and 'homozyg-kb' parameters in order to call ROHs with a minimum of 30, 40 and 50 SNPs or 2, 4, and 10Mb, as indicated in Table 6.1.

.

Owing to the lower density of SNPs in the low-LD panel compared to the 500K SNP panel, the parameters for minimum density of SNPs (homozyg-density) and maximum distance between SNPs (homozyg-gap) were altered to remove these limitations. For the 'homozyg-density' parameter, a value of 50 specifies that there must be 1 SNP in every 50Kb. If two SNPs within a segment are too far apart, as defined by homozyg-gap, then the segment will be split in two. The parameters set for the less dense low-LD panel prevent this occurring over small runs of continuous homozygous SNPs.

Table 6.1 The ROH calling parameters

The parameters used in calling a run of homozygous SNPs as an ROH using the homozyg function in plink for the 500K and Low-LD SNP panels.

ROH size criteria	ROH calling parameters					
	homozyg-snp	homozyg-kb	homozyg-gap (kb)		homozyg-density (kb/SNP)	
			Low-LD	500K	Low-LD	500K
≥30 SNPs	30	0.01	1,000,000	1000	1,000,000	50
≥40 SNPs	40	0.01	1,000,000	1000	1,000,000	50
≥50 SNPs	50	0.01	1,000,000	1000	1,000,000	50
≥2Mb	1	2000	1,000,000	1000	1,000,000	50
≥4Mb	1	4000	1,000,000	1000	1,000,000	50
≥10Mb	1	10000	1,000,000	1000	1,000,000	50

6.2.4 Statistical Analysis

The association of increased homozygosity with CRC was tested using two different approaches. The first was to test homozygote frequency by SNP and the second was an analysis of the size and number of ROHs detected in cases and controls. As PLINK did not have the flexibility to undertake variations on the basic association analyses already available within the program, all statistical analyses were performed using my own scripts for packages available in R. The full details for each of the main analyses are given in the Materials and Methods, Section 2.11.

6.2.5 Imputation of SNPs not genotyped in VQ58

The imputation of SNPs not genotyped in VQ58, but present on the HumanHap550, was performed using IMPUTE (v0.5) and the build 35 HapMap Phase II reference panel. The imputed SNP genotype probabilities were converted into ‘best guess’ genotypes, using a threshold probability of 0.9, and incorporated into the meta-analysis with the genotyped EngP1 data. No SNPs were imputed in the controls in VQ58, as they were genotyped on the HumanHap550. Any imputed SNPs that did not achieve an information score greater than 0.5 and an overall call rate of 90% were removed from the analysis.

6.2.6 The Inbreeding Coefficient

The inbreeding coefficient (F) was introduced in 1922 to quantify genetic relatedness (Wright 1922). F is a measure of the difference between the frequency of heterozygotes in an individual and the expected frequency of heterozygotes when genotypes are in HWE. In the presence of inbreeding, or fewer heterozygotes than expected, F will be positive. However, if F is negative then the individual has more heterozygotes than expected under HWE and is not inbred (Holsinger and Weir 2009). A strongly negative F statistic value can indicate sample contamination. Therefore, the F statistic should be close to zero for outbred individuals. PLINK contains a function (het) to calculate the F statistic for each individual based on this coefficient and is calculated as follows. If p and q represent the allele frequencies of a SNP then the probability of homozygosity at this SNP in individual *i*, is the probability of being autozygous (f_i) plus the probability of being homozygous by chance.

$$Prob(i)_{homoz} = f_i + (1 + f_i)(p^2 + q^2)$$

The F statistic is then be calculated by:

$$f_i = \frac{(Obs_i - Exp_i)}{(L_i - Exp_i)}$$

Where L_i is the number of genotyped autosomal SNPs, Obs_i is the number of observed heterozygotes and Exp_i is the number of expected heterozygotes under HWE. In the absence of known allele frequencies, the expected number of heterozygotes is based on the sum for all SNPs observed in the individual (Purcell *et al.* 2007).

6.3 Results

6.3.1 GWA association analysis between homozygosity and CRC

To test whether being homozygous at an individual SNP, regardless of allele, is associated with CRC risk, I performed an association analysis, as described in the Materials and Methods, using the 500K SNP panel in EngP1. No SNPs achieved a globally significant P value, but 35 SNPs achieved a P_{homoz} of less than 1x10^{-4} (results are given in Table 6.2). The most strongly associated SNP, rs17062732, was located on chromosome 13 at 41,495,284bp (P_{homoz}=5.90x10^{-06}, OR=1.61). There are no known genes in close proximity to this SNP.

Table 6.2 Results for the association analysis between homozygosity and CRC in the EngP1 dataset

The table shows the SNPs where P_{homoz} values were less than 1x10^{-4} for EngP1 using the 500K SNP panel. The positions of the SNPs listed are the genome build 35 positions.

SNP	Chr.	Position	P_{homoz}	OR	AABB aff	AB aff	AABB ctrl	AB ctrl
rs17062732	13	41,495,284	5.90E-06	1.611	705	216	622	307
rs9293478	5	86,130,471	1.42E-05	0.657	514	407	611	318
rs1860345	12	4,781,073	1.85E-05	1.507	586	335	499	430
rs6029910	20	40,042,498	2.09E-05	0.666	487	434	583	346
rs2434137	4	138,951,056	2.32E-05	0.644	606	313	697	232
rs1884033	20	40,117,008	2.51E-05	0.662	539	381	633	296
rs8102662	19	59,543,040	4.33E-05	0.680	426	495	519	410
rs6012416	20	46,467,236	4.45E-05	0.680	405	512	497	427
rs11076194	16	50,460,971	4.66E-05	0.681	411	509	504	425
rs12572686	10	16,759,094	5.08E-05	1.466	517	404	433	496
rs2839657	10	31,695,774	5.23E-05	0.683	417	504	509	420
rs8008317	14	85,203,204	5.27E-05	1.465	499	422	415	514
rs7958635	12	112,951,388	5.31E-05	0.683	443	478	535	394
rs2215439	7	84,440,361	5.33E-05	0.680	482	439	573	355
rs12754637	1	4,933,694	5.33E-05	1.516	682	239	606	322
rs1884040	20	40,139,908	5.80E-05	0.681	491	430	582	347
rs618236	18	38,588,098	6.34E-05	1.459	506	415	423	506
rs1348271	11	99,642,556	6.43E-05	0.676	509	394	596	312
rs9477166	6	16,673,171	6.90E-05	0.623	701	220	777	152
rs11075365	16	17,567,136	7.06E-05	0.687	423	497	514	415
rs4755201	11	33,558,223	7.31E-05	1.470	610	311	531	398
rs4755718	11	33,559,489	7.32E-05	1.472	610	309	527	393
rs1840819	19	33,425,918	7.37E-05	0.630	692	229	768	160
rs1904833	5	105,058,907	7.47E-05	0.660	612	302	697	227
rs10089677	8	122,729,429	7.80E-05	1.452	498	423	416	513

rs2888998	9	30,969,640	7.91E-05	1.626	764	136	715	207
rs1864125	5	154,549,343	8.03E-05	0.664	610	311	694	235
rs2019083	2	221,853,171	8.11E-05	1.452	518	402	434	489
rs12037907	1	67,684,456	8.14E-05	0.620	712	209	786	143
rs4597574	2	30,916,948	8.68E-05	0.690	439	481	529	400
rs11636893	15	91,709,406	8.95E-05	1.506	673	229	611	313
rs1557800	8	31,553,927	9.13E-05	1.518	709	212	639	290
rs1325619	13	78,026,569	9.45E-05	0.691	457	464	546	383
rs704409	3	64,221,869	9.48E-05	1.445	496	425	415	514
rs4806074	19	33,403,638	1.00E-04	0.651	659	262	738	191

As a validation stage, the results of all SNPs with P_{homoz} of less than 1x10^{-4} (35 SNPs) were then combined in a meta-analysis with the VQ58 dataset using both fixed and random effects models (meta-analysis was performed using a script in R, which is given in the materials and methods and the results are given in Table 6.3). The most associated SNP in the original analysis of EngP1, rs17062732, was not replicated in VQ58 and showed no evidence of an association with CRC (P_{homoz}=0.0038, OR=1.2, fixed effects model).

The most significant SNP from the overall meta-analysis of EngP1 and VQ58 was rs6029910 (chr20:40,042,498, P_{homoz}=2.86x10^{-05}, OR=0.776, fixed effects model). However, this SNP had a P value of 0.037 under a random effects model and showed evidence of heterogeneity between studies.

Table 6.3 Results table for the meta-analysis of EngP1 and VQ58

The results of the meta-analysis showing the most strongly associated 35 SNPs from the original analysis in EngP1. The SNP with the lowest P value in the EngP1 analysis is highlighted in bold.

SNP	Chr.	Position (bp)	Genotyped/ Imputed in VQ58	Fixed effects P_{homoz}	Fixed effects OR	Random Effects P_{homoz}	Random Effects OR
rs6029910	20	40,042,498	Genotyped	2.86x10^{-5}	0.776	0.037	0.762
rs6012416	20	46,467,236	Genotyped	7.87x10^{-5}	0.789	0.045	0.775
rs4806074	18	33,403,638	Genotyped	3.04x10^{-4}	0.775	0.07	0.761
rs9293478	5	86,130,471	Imputed	5.54x10^{-4}	0.809	0.176	0.787
rs1884040	20	40,139,908	Genotyped	5.85x10^{-4}	0.812	0.124	0.794
rs1840819	19	33,425,918	Genotyped	1.29x10^{-3}	0.790	0.165	0.767
rs1884033	20	40,117,008	Genotyped	1.39x10^{-3}	0.820	0.213	0.797
rs2888998	9	30,969,640	Genotyped	1.65x10^{-3}	1.273	0.177	1.318

rs12572686	10	16,759,094	Genotyped	1.70x10^{-3}	1.207	0.198	1.237
rs1864125	5	154,549,343	Genotyped	1.83x10^{-3}	0.814	0.187	0.794
rs17062732	**13**	**41,495,284**	**Imputed**	**3.81x10^{-3}**	**1.216**	**0.345**	**1.259**
rs8008317	14	85,203,204	Genotyped	5.53x10^{-3}	1.180	0.287	1.215
rs4755201	11	33,558,223	Genotyped	6.31x10^{-3}	1.185	0.291	1.217
rs1325619	13	78,026,569	Genotyped	7.37x10^{-3}	0.852	0.288	0.828
rs4755718	11	33,559,489	Imputed	8.81x10^{-3}	1.177	0.321	1.210
rs10089677	8	122,729,429	Genotyped	0.010	1.167	0.322	1.202
rs2215439	7	84,440,361	Genotyped	0.011	0.858	0.345	0.831
rs704409	3	64,221,869	Genotyped	0.014	1.157	0.351	1.193
rs1904833	5	105,058,907	Genotyped	0.016	0.852	0.372	0.824
rs2019083	2	221,853,171	Genotyped	0.017	1.154	0.376	1.190
rs11075365	16	17,567,136	Genotyped	0.026	0.875	0.420	0.847
rs7958635	12	112,951,388	Genotyped	0.032	0.880	0.448	0.849
rs4597574	2	30,916,948	Imputed	0.039	0.883	0.456	0.855
rs2434137	4	138,951,056	Genotyped	0.043	0.876	0.493	0.837
rs2839657	10	31,695,774	Genotyped	0.046	0.888	0.485	0.856
rs11076194	16	50,460,971	Genotyped	0.066	0.896	0.524	0.862
rs1348271	11	99,642,556	Imputed	0.073	0.893	0.551	0.865
rs12037907	1	67,684,456	Imputed	0.109	0.885	0.572	0.843
rs1557800	8	31,553,927	Genotyped	0.115	1.110	0.555	1.166
rs1860345	12	4,781,073	Genotyped	0.120	1.099	0.609	1.148
rs618236	18	38,588,098	Genotyped	0.185	1.082	0.632	1.129
rs9477166	6	16,673,171	Imputed	0.569	0.954	0.939	0.967

Ten of these SNPs were not directly genotyped in VQ58 and were imputed, including the top SNP from EngP1, rs17062732. Three of these SNPs (rs8102662, rs12754637 and rs11636893) failed the imputation process and were removed from the analysis. A summary of the quality metrics for the imputed SNPs is given in Table 6.4.

Table 6.4 The quality scores for the imputed SNPs

The table shows summary information for the ten SNPs that were imputed in the VQ58 cases to facilitate meta-analysis with the EngP1 results. rs12754637 and rs8102662 were excluded owing to info scores below 0.5 and rs11636893 was excluded owing to the low call rate in the imputed cases.

SNP	Chr.	Position (bp)	IMPUTE Info score	Call rate aff	Call rate Ctrl	MAF Aff	MAF Ctrl
rs12037907	1	67,684,456	0.935	0.98	1.000	0.185	0.111
rs12754637	1	4,933,694	0.243	0.97	0.998	0.001	0.201
rs4597574	2	30,916,948	0.983	0.998	0.999	0.295	0.382
rs9293478	5	86,130,471	0.982	0.976	1.000	0.272	0.280
rs9477166	6	16,673,171	0.784	0.994	0.995	0.241	0.087
rs4755718	11	33,559,489	0.994	1	0.992	0.224	0.220
rs1348271	11	99,642,556	0.920	0.902	0.970	0.699	0.264
rs17062732	13	41,495,284	0.963	0.922	0.999	0.862	0.166
rs11636893	15	91,709,406	0.743	0.719	0.984	0.094	0.177

rs8102662	19	59,543,040	0.316	0.011	1.000	0.500	0.427

6.3.1.1 Recessive association tests

An increase in homozygosity in an individual is suggested to lead to decreased health and overall fitness owing to the increase in homozygous recessive deleterious alleles across the genome. Therefore, to test whether being homozygous for the minor allele at a given SNP was associated with CRC risk, I re-analysed the data using the recessive models in PLINK. No SNPs reached global significance. The most significant results are given in Table 6.5.

Table 6.5 Most significant results for the recessive tests in EngP1

This table provides the results of an association analysis using the recessive model for disease inheritance. This analysis tests whether being homozygous for the minor allele, compared with being heterozygous or homozygous for the major allele is associated with disease risk (AA vs AB,BB)

Test	Chr.	SNP	Position (bases, build35)	Alleles (minor/major)	P value	OR
Recessive	22	rs2073989	19,669,438	T/C	1.43x10^{-5}	1.86
Recessive	4	rs7676572	189,007,130	G/A	1.98x10^{-5}	0.61

6.3.1.2 Analysis of overall level of homozygosity across all SNPs

The results above demonstrate that, in EngP1, no individual SNP is overrepresented as a homozygote in cases compared with controls. In order to determine whether cases had more homozygous genotypes overall than controls, I simply counted homozygous genotypes, per individual, in both the 500K and low-LD SNP panels, generated summary statistics and then performed a non-parametric Wilcoxon rank sum test. The mean number of homozygous genotypes using the 500K SNP panel was 315,644 (median=316,779; SD= 2,059) in the cases and 315,583 (median=315,615; SD=1,781) in the controls. The difference between cases and controls was statistically significant (P=0.029, Wilcoxon test). However, when I repeated this analysis in the low-LD panel the difference in the mean of homozygous genotypes between cases and controls was negligible. The mean in cases was 21,499 (median=21,506, SD=126) and controls 21,496 (median=21,501, SD=111) and was not significant (P=0.397, Wilcoxon test).

I then applied the F statistic (inbreeding coefficient) to all samples in the study using the low-LD panel of SNPs in order to identify any individuals that were likely to be the offspring of consanguineous relationships and enable comparison with the total size of ROHs. The mean F statistic in the controls was 0.00101 (SD=0.012) and in the cases was 0.00135 (SD=0.012). There was no evidence to suggest that cases were more inbred than controls (P=0.54, t test).

6.3.2 Analysis of ROHs

I then compared the total size and total number of ROHs in cases and controls to determine if the presence of continuous runs of homozygous genotypes were associated with increased CRC risk. The ‘homozyg’ function produces several output files. The plink.hom file provides a breakdown of every ROH detected in every sample and includes the sample ID, phenotype, start and end positions for the ROH, size and number of SNPs covered by the ROH. This file was used for visualising the locations of the ROHs detected. The plink.hom.indiv output file provides a summary of the ROHs detected and gives the total number and total size of all ROHs detected in each individual analysed. The data from this file were used in the comparison of ROHs detected in cases and controls.

6.3.2.1 Comparison of the number of ROHs

In the study by Bacalod *et al.*, the authors called a run of homozygous SNPs as an ROH if it included more than 50 SNPs and covered more than 4Mb. Using this criteria, they discovered ROHs in 62.2% of cases and in 35.6% and 28.8% in the two control groups. In order to allow a comparison, I calculated the frequencies of samples with an ROH larger than 4Mb, but failed to detect a significant difference between cases and controls. Using the 500K panel, there were 159 out of 921 (17%) cases and 142 out of 929 (15%) controls with ROHs, but this slight difference was not significant (P=0.14, Fisher's exact test). When this analysis was repeated in the low-LD panel, 8 out of 921 cases (0.87%) and 8 out of 929 controls (0.86%) had ROHs (P=0.59, Fisher's exact test).

6.3.2.2 Analysis of total size of ROH

In order to provide a better representation of the overall level of homozygosity in each individual, I then used the total length (the sum of all the lengths of ROHs) rather than the total number of ROHs detected. If the total number of ROHs was used, an individual that has four 2Mb ROHs would be scored with more weight than an individual with two 4Mb ROHs when, in fact, they are equally important. Therefore, total ROH length should be more appropriate for illustrating autozygosity or indicating relative measures of overall homozygosity than total counts of ROHs.

We did not have a clear idea of the most informative method for ROH analysis and, therefore, the analysis was performed using a number of different criteria for calling an ROH to determine whether the method was sound. ROHs were defined by more than or equal to 30 SNPs, 40 SNPs, 50 SNPs, 60 SNPs, 2Mb, 4Mb or 10Mb and the analysis repeated in both the 500K and the low-LD SNP panels (for summaries of the total ROH size data per individual in cases and controls using both SNP panels see Table 6.6).

Using the 500K SNP panel, I detected at least 100 ROHs (the mean number of ROHs per individual was 494.2), covering more than 50 SNPs, in every individual in the study. However, when the total ROH size for each individual was compared in cases versus controls, I found no evidence to suggest an association between CRC and total ROH size (P=0.29, Wilcoxon test, see Table 6.6). The same conclusions were drawn when each of the various calling criteria were used.

When I repeated this analysis using the low-LD SNP panel, far fewer ROHs were detected, with the majority of individuals having no ROHs detectable using the criteria above. Those that were detected were large, spanning more than 2Mb in length on average. However, as with the 500K analysis, the difference between cases and controls was not significant.

Table 6.6 The total size of ROHs detected in cases and controls

Table A) shows the 500K and B) the Low-LD SNP panel. ROHs were defined using the size or number of SNPs criteria shown in the first column and the total ROH size calculated as a sum of all ROHs meeting the defining criteria in each individual. The Wilcoxon rank sum test was performed by comparing the total ROH size between cases and controls (including 929 controls and 921 cases). The mean size per ROH is for cases and controls combined. Total ROH size is bigger in the 500K panel owing to the much higher density of SNPs.

A) The 500K SNP panel

Min. ROH Size	Phenotype	Total ROH size summaries				Mean size per ROH (kb)	P (Wilcoxon)
		Mean (kb)	Median (kb)	Range			
				Min (kb)	Max (kb)		
≥30 SNPs	Cases Controls	235,920 235,204	236,368 235,533	49,364 21,725	546,053 426,693	437	0.353
≥40 SNPs	Cases Controls	235,589 234,874	235,938 235,299	48,925 21,725	546,053 426,693	436	0.354
≥50 SNPs	Cases Controls	222,949 222,113	222,863 222,020	40,770 19,181	531,834 413,025	496	0.269
≥60 SNPs	Cases Controls	167,096 166,248	166,806 166,366	23,326 12,006	487,598 366,155	528	0.286
≥2Mb	Cases Controls	7,024 6,819	5,210 5,296	0 0	334,589 221,843	2,987	0.669
≥4Mb	Cases Controls	1,785 1,666	0 0	0 0	317,328 216,989	7,291	0.323
≥10Mb	Cases Controls	656 620	0 0	0 0	260,696 204,059	19,674	0.984

B) The Low-LD SNP panel

Min. ROH Size	Phenotype	Total ROH size summaries				Mean size per ROH (kb)	P (Wilcoxon)
		Mean (kb)	Median (kb)	Range			
				Min (kb)	Max (kb)		
≥30 SNPs	Cases Controls	1,367 1,291	0 0	0 0	354,933 225,157	9,954	0.3177
≥40 SNPs	Cases Controls	1,367 1,291	0 0	0 0	354,933 225,157	9,954	0.3177
≥50 SNPs	Cases Controls	1,322 1,230	0 0	0 0	354,933 225,157	10,399	0.253
≥60 SNPs	Cases Controls	1,145 1,125	0 0	0 0	352,932 225,157	12,205	0.4152
≥2Mb	Cases Controls	1,350 1,279	0 0	0 0	354,933 225,157	10,526	0.4928
≥4Mb	Cases Controls	1,274 1,214	0 0	0 0	346,623 225,157	12,243	0.4668
≥10Mb	Cases Controls	807 807	0 0	0 0	304,718 219,711	24,075	0.9834

Owing to the difference in SNP density between the 500K and the low-LD panel (30K SNPs), 30 SNPs in the low LD panel covers a much larger distance than the same number of SNPs in the 500K panel. For this reason, the mean ROH size in the low LD panel, especially for the ROHs defined by number of SNPs, is substantially greater. The difference in the mean ROH size for the ROHs that were defined by size is probably due to the presence of LD between SNPs extending the ROH in the 500K panel.

It may seem interesting that in the table above for the 500K panel analysis the controls appear to have samples with smaller ROH sizes than the cases, as seen in the range by the minimum total ROH size and that the maximum size is always larger in cases than in controls. For example, for the more than 50 SNP group, cases have a minimum total size of 40,770kb while controls have a minimum total size of 19,181kb. However, it should be noted that there was only one affected individual with a total ROH size greater than the maximum seen in controls (sample 1053H10) and there was also only one control sample with a total ROH size less than the minimum seen in cases (sample 1060C04) as can be seen in Figure 6.1.

Figure 6.1 The total ROH size distribution detected in the 500K and LowLD SNP panels

Cases are shaded green and controls are white with diagonal stripes. The total ROH size per person is plotted against the number of samples (n) on a logarithmic scale. A count of 1 has been added to each total count to enable the counts of 1 to be visible in the plot. Data shown is for ROHs >50 SNPs and >4Mb. Cases are shaded in black and controls are in grey.

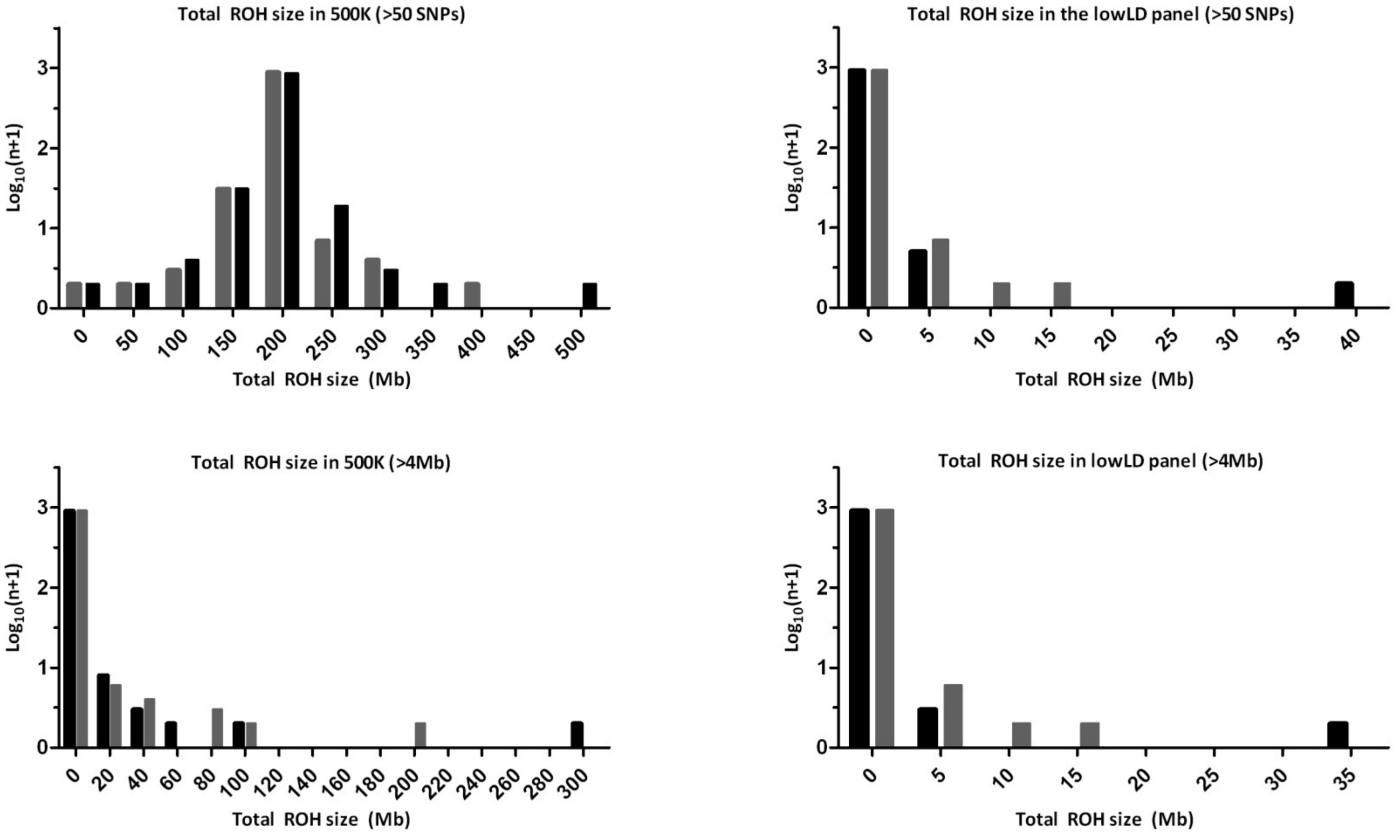
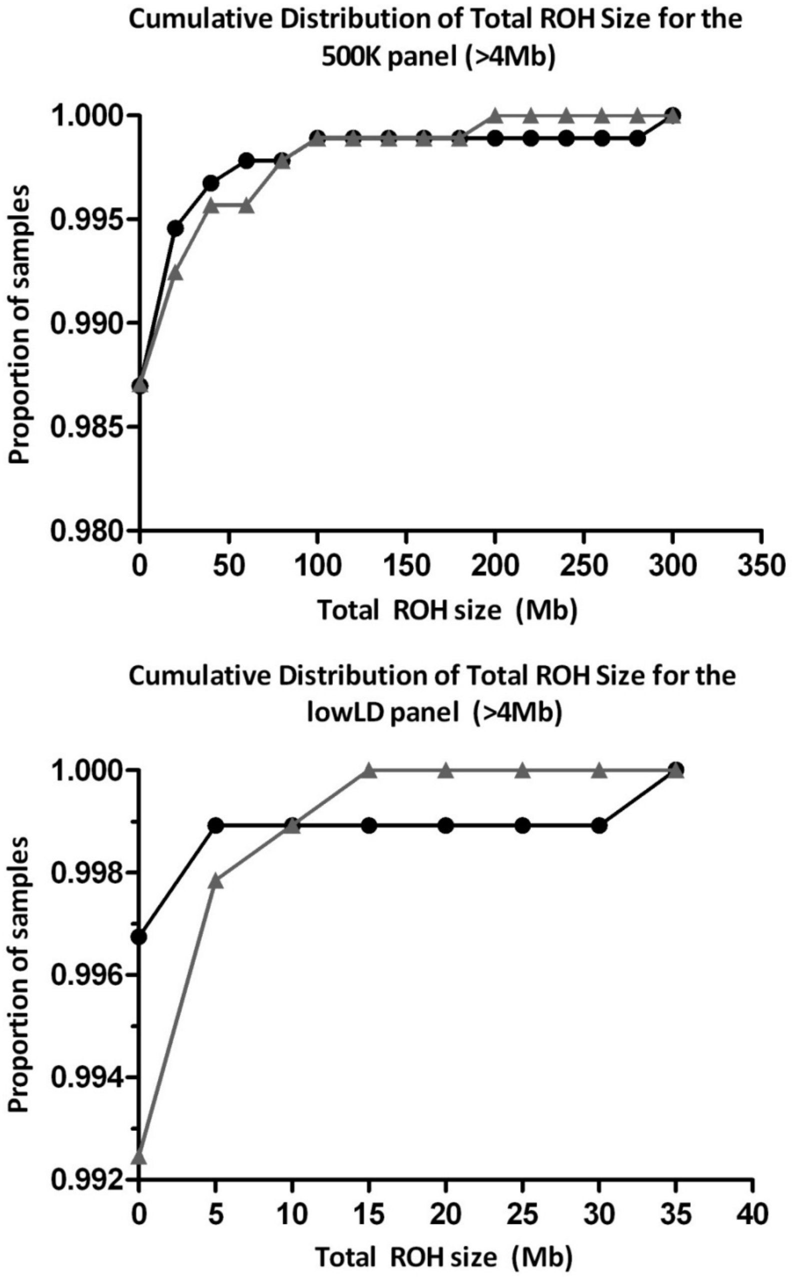


Figure 6.2 Cumulative distribution for the total ROH size against the proportion of samples for ROHs >4Mb

The data for both the 500K and the LowLD panels are shown. The cases are plotted in black and the controls in grey.



Cumulative Distribution of Total ROH Size for the 500K panel (>4Mb)

Cumulative Distribution of Total ROH Size for the lowLD panel (>4Mb)

The chromosomal positions of the ROHs detected in all samples that were defined using the 50 SNP and 4Mb criteria for the low-LD SNP panel are shown in Figure 6.3

and Figure 6.4, respectively. The five longest ROHs belong to just two individuals (the case: 1053H10 and the control: 1049G06). These samples were also evident as outliers in the total ROH size distribution graphs, for the 500K and low-LD SNP panels, in Figure 6.1 above. The inbreeding coefficient (F) for both these samples is higher than average (F=0.1, total ROH size=354,933kb for 1053H10 and F=0.066, total ROH size= 225,157kb for 1049G06), although the pedigrees from these individuals showed no evidence of consanguineous relationships. Sample 1053H10 had an unremarkable family history and was diagnosed with a Dukes A left sided CRC at age 60 years.

As expected, a linear regression analysis of the total ROH size and the F statistic did show a direct correlation ($P<2.2x10^{-16}$, calculated using ROHs greater than 4Mb in size). However, this was not significant when comparing cases against controls. The full data table showing these data for each sample was provided in the online supplementary material for the paper where these results were published (Spain *et al.* 2009).

For the 500K panel, owing to the large number of ROHs detected, I have only included the plot showing ROHs greater than 4Mb (see Figure 6.5). The genome-wide ROH plot shows a reasonably high frequency of ROHs at chromosome 11p11 and 6p22.1. These regions correspond with those of long-range LD at chr6:25.5-33.5Mb and chr11:46-57Mb that were identified in European populations (Price *et al.* 2008) by PCA analysis. The results also indicate that uniparental isodisomy, as a cause of ROHs, is not likely to be a common feature of these samples as ROHs covering telomeres, whole chromosomes or chromosome arms were not seen.

Figure 6.3 The genome-wide view plot for the low-LD panel showing ROH >50 SNPs

These plots were created using a command line version of the Genome-wide viewer program developed by Jean-Baptiste Cazier (http://www.well.ox.ac.uk/~jcazier/GWA_View.html).

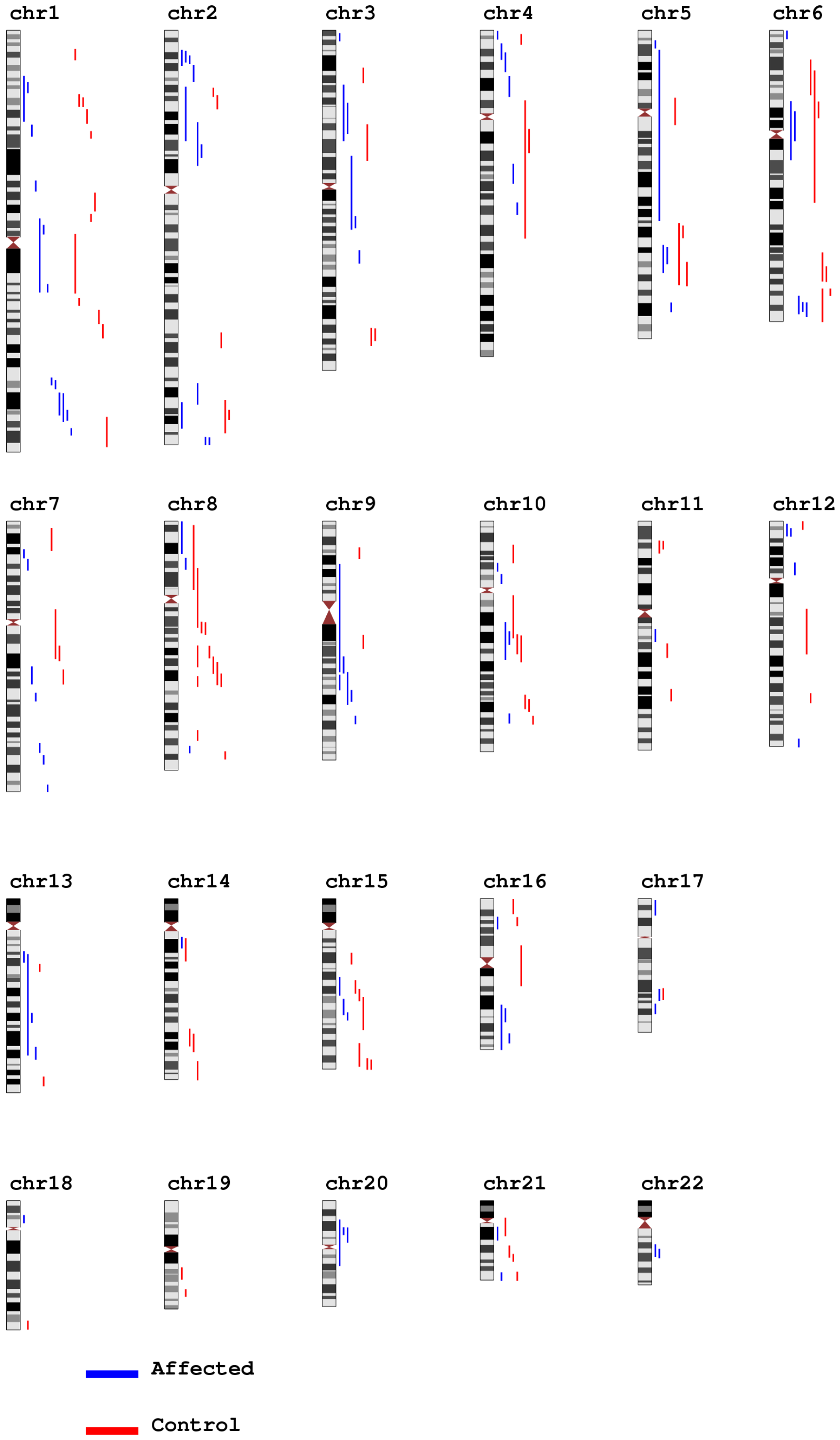


Figure 6.4 The genome-wide view plot for the low-LD panel showing ROHs >4Mb

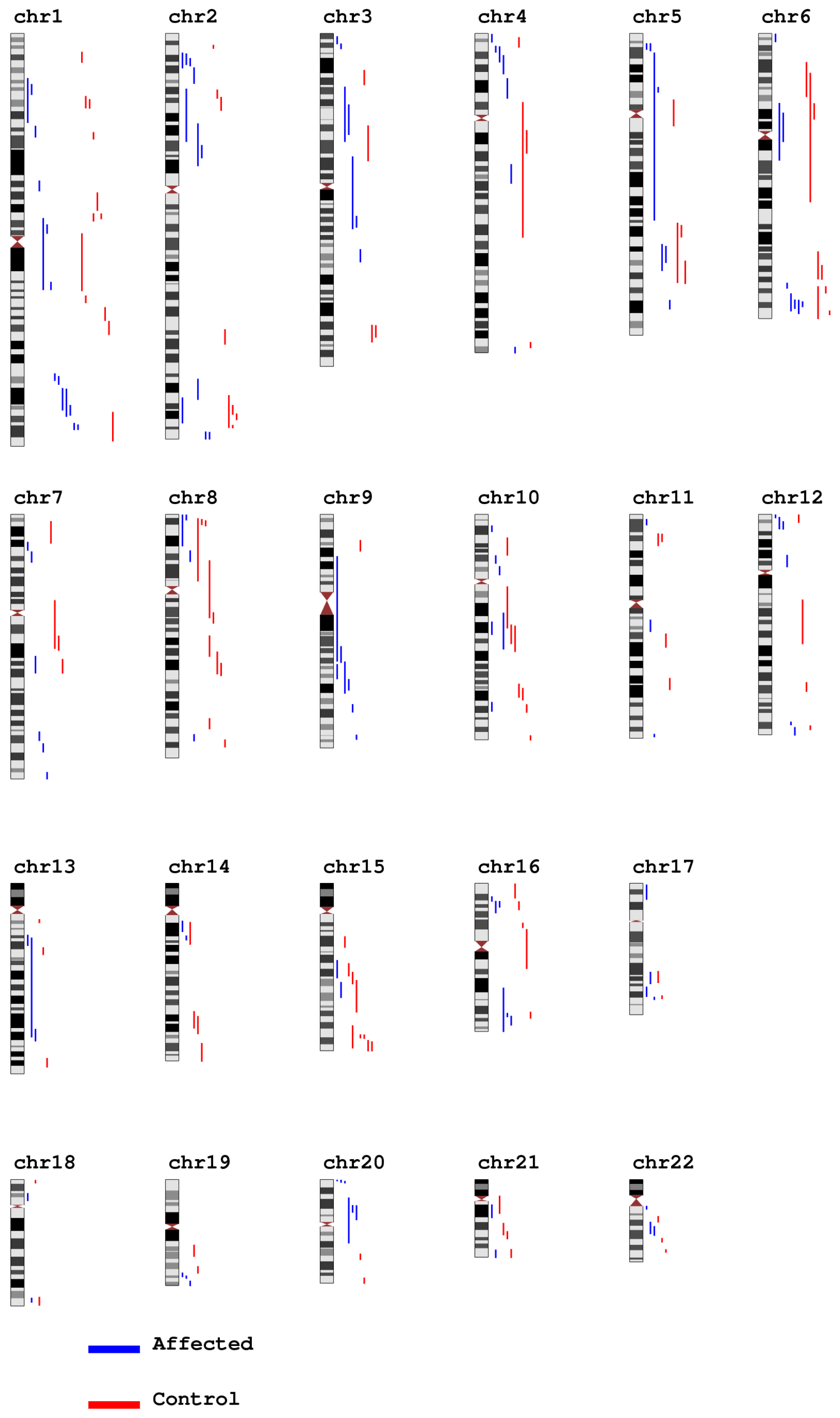
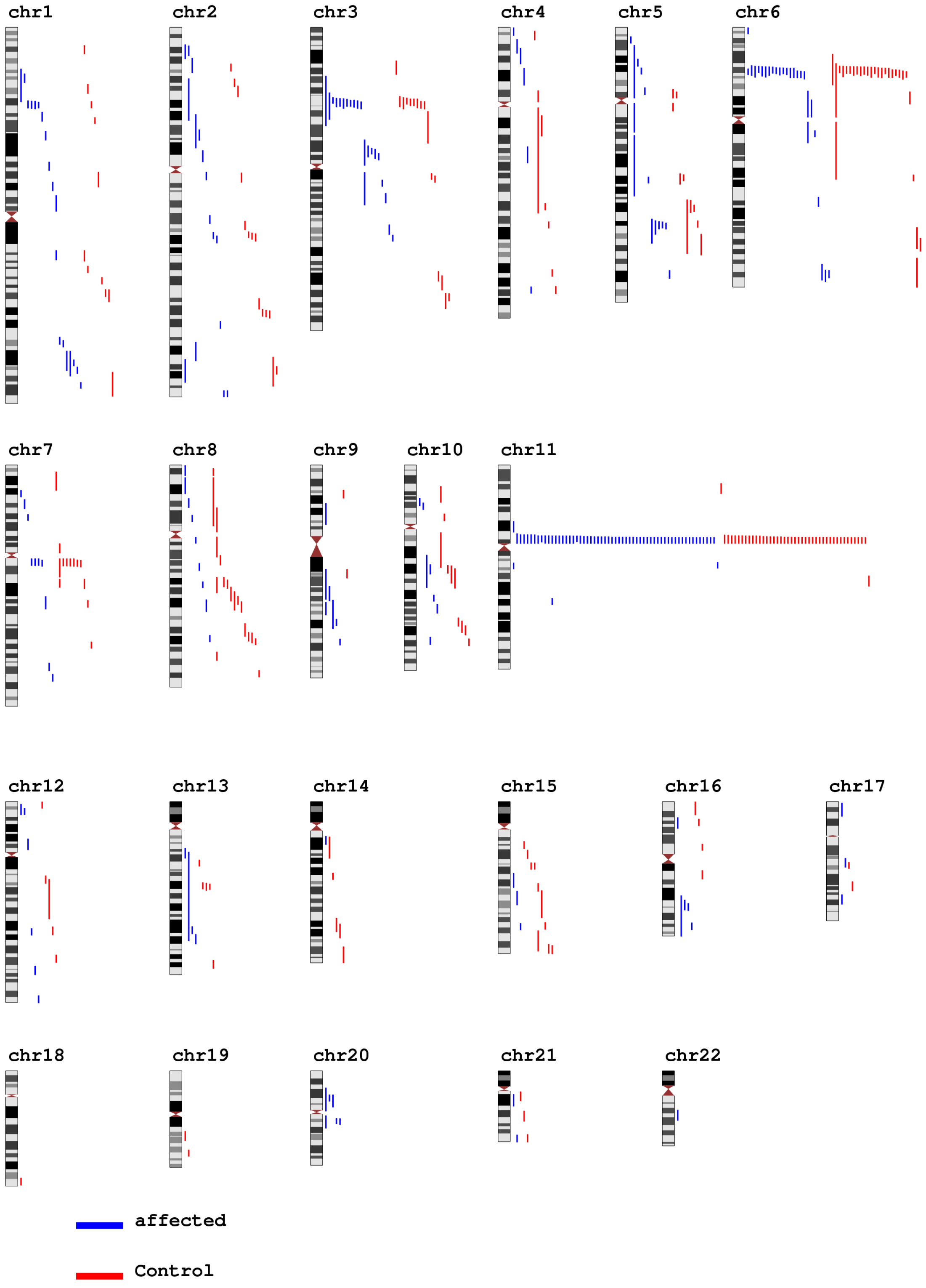


Figure 6.5 The genome-wide view plot for the 500K panel showing ROHs >4Mb



6.3.2.3 Common regions of ROHs and association with CRC

So far I have only examined ROHs across all samples where the actual regions of individual ROH were quite uncommon. However, using the 500K panel, I did detect relatively short ROHs that were present in more than 10% of the samples. I therefore, extended the analysis to study whether any of these recurrent regions were more common in cases compared to controls and associated with CRC risk.

A common, or recurrent, region was defined by the occurrence of at least 5 ROHs, greater than 1Mb, that overlap to give a consensus region. I only considered consensus regions that spanned more than one SNP and then compared the number of cases and controls with an ROH overlapping the consensus region to determine whether there was any association of the common regions with CRC. I initially analysed the 500K panel data and only included common regions that were seen in more than five samples (cases and/or controls). 3,478 common regions that met the above criteria were identified by searching for minimal overlapping regions between all detected ROHs. However, none of these homozygous regions were significantly associated with CRC risk after multiple testing was taken into account (see Table 6.7). The three most associated regions, which were more common in controls compared to cases, were located on chromosome 2 at approximately 160Mb, P=1.62×10^{-4} (OR_{homoz}=0.164).

Table 6.7 The recurrent ROH regions in the 500K SNP panel

The common ROH regions that achieved a P value less than 0.05, which were identified in more than 5 individuals and with a consensus region spanning more than 1 SNP are listed here. The start and end positions given provide the size and location of the consensus region. P values were calculated using a χ^2 test of the counts in cases and controls. Where the count was below 5 a Fisher's test was used instead. Only ROHs greater than 1Mb were used to determine the common regions.

ROH	Chr	Start Position	End Position	Size (kb)	P_{homoz}	OR	ROH aff	No ROH aff	ROH ctrl	No ROH ctrl
S18265	2	160,511,276	160,514,064	2.79	1.62x20^{-4}	0.164	4	917	24	905
S17452	2	160,840,573	160,904,942	64.37	2.93x10^{-4}	0.197	5	916	25	904
S17453	2	160,606,433	160,705,158	98.73	2.93x10^{-4}	0.197	5	916	25	904
S29628	4	77,456,923	77,501,861	44.94	0.002	0.000	0	921	10	919
S29629	4	77,196,908	77,338,385	141.48	0.002	0.000	0	921	10	919
S2080	8	112,473,978	112,508,791	34.81	0.002	0.621	79	842	122	807
S30874	20	17,861,286	18,857,730	996.44	0.004	Inf	8	913	0	929
S31138	10	95,008,372	95,064,402	56.03	0.004	Inf	8	913	0	929
S31676	2	108,344,413	108,384,090	39.68	0.004	Inf	8	913	0	929
S20059	3	104,449,957	104,463,134	13.18	0.004	3.893	19	902	5	924
S13683	16	67,942,360	67,959,438	17.08	0.004	2.810	30	891	11	918
S23143	13	80,660,902	80,693,350	32.45	0.004	5.110	15	906	3	926
S2133	8	49,720,677	49,744,192	23.52	0.004	0.639	79	842	119	810
S1623	8	112,226,716	112,295,736	69.02	0.005	0.669	97	824	139	790
S1624	8	112,101,273	112,167,469	66.20	0.005	0.669	97	824	139	790
S13947	16	67,745,613	67,867,637	122.02	0.006	2.713	29	892	11	918
S13948	16	67,648,719	67,728,069	79.35	0.006	2.713	29	892	11	918
S26225	6	110,414,498	110,425,208	10.71	0.007	6.119	12	909	2	927
S23801	13	80,508,388	80,547,295	38.91	0.007	4.764	14	907	3	926
S32568	4	27,501,432	28,213,043	711.61	0.007	Inf	7	914	0	929
S22031	10	31,519,109	31,842,788	323.68	0.011	0.249	4	917	16	913
S27286	2	85,934,680	85,956,820	22.14	0.012	5.603	11	910	2	927
S15316	3	166,084,977	166,141,326	56.35	0.012	0.381	10	911	26	903
S6434	8	113,734,730	113,742,508	7.78	0.013	0.572	35	886	60	869
S5202	8	113,554,111	113,555,392	1.28	0.013	0.601	43	878	70	859
S18214	3	166,354,282	166,372,522	18.24	0.014	0.331	7	914	21	908
S19457	6	110,338,744	110,346,363	7.62	0.015	3.240	19	902	6	923
S19578	3	102,093,722	102,191,686	97.96	0.015	3.240	19	902	6	923
S32379	6	136,546,729	136,692,187	145.46	0.015	0.000	0	921	7	922
S20883	10	31,974,003	32,027,252	53.25	0.016	0.293	5	916	17	912
S19615	2	161,194,412	161,200,547	6.14	0.017	0.314	6	915	19	910
S5264	8	113,498,812	113,510,014	11.20	0.017	0.610	43	878	69	860
S9966	2	131,636,064	131,646,470	10.41	0.017	1.963	40	881	21	908
S23121	14	67,200,289	67,318,698	118.41	0.018	3.569	14	907	4	925
S23142	13	81,019,068	81,035,524	16.46	0.018	3.569	14	907	4	925
S15034	3	112,534,749	112,559,738	24.99	0.019	2.424	26	895	11	918
S5004	8	113,452,239	113,456,509	4.27	0.019	0.621	45	876	71	858
S16710	2	57,996,894	58,393,514	396.62	0.019	2.618	23	898	9	920
S8634	2	131,475,203	131,478,698	3.50	0.020	1.826	46	875	26	903
S30240	9	69,198,209	69,503,452	305.24	0.021	8.131	8	913	1	928

S29541	5	113,239,111	113,282,656	43.55	0.021	0.111	1	920	9	920
S18479	10	32,154,956	32,230,618	75.66	0.021	0.348	7	914	20	909
S11485	10	116,642,492	116,648,096	5.60	0.021	2.056	34	887	17	912
S27581	11	110,659,867	110,663,399	3.53	0.022	5.088	10	911	2	927
S27712	8	55,329,229	55,426,965	97.74	0.022	5.088	10	911	2	927
S27713	8	54,917,377	55,217,668	300.29	0.022	5.088	10	911	2	927
S6604	8	113,756,657	113,759,975	3.32	0.022	0.593	35	886	58	871
S15813	9	128,390,682	128,716,989	326.31	0.023	2.459	24	897	10	919
S15814	9	128,168,018	128,225,963	57.95	0.023	2.459	24	897	10	919
S3401	13	56,271,383	56,289,805	18.42	0.024	1.496	89	832	62	867
S13817	4	171,603,813	171,611,973	8.16	0.025	2.209	28	893	13	916
S3235	13	56,091,802	56,108,951	17.15	0.025	1.482	91	830	64	865
S21970	11	111,497,802	111,536,290	38.49	0.025	3.060	15	906	5	924
S21981	11	64,341,563	64,356,578	15.02	0.025	3.060	15	906	5	924
S1420	11	47,813,829	47,830,459	16.63	0.026	1.366	144	777	111	818
S21437	10	47,155,834	47,166,092	10.26	0.026	0.311	5	916	16	913
S21598	5	15,288,921	15,601,576	312.66	0.026	0.311	5	916	16	913
S15895	6	65,589,787	65,648,275	58.49	0.026	0.414	10	911	24	905
S3284	8	112,584,436	112,617,527	33.09	0.027	0.676	63	858	91	838
S8421	4	53,484,407	53,550,105	65.70	0.027	0.567	27	894	47	882
S13209	12	79,854,307	79,977,204	122.90	0.029	2.125	29	892	14	915
S23859	11	64,550,536	64,575,744	25.21	0.030	3.311	13	908	4	925
S4071	10	22,135,743	22,179,540	43.80	0.030	0.665	55	866	81	848
S1491	8	49,401,130	49,415,329	14.20	0.030	0.736	108	813	142	787
S14682	4	171,302,775	171,319,428	16.65	0.031	2.220	26	895	12	917
S10843	5	93,076,209	93,465,676	389.47	0.031	0.523	19	902	36	893
S33509	6	15,393,634	16,206,969	813.34	0.031	0.000	0	921	6	923
S8750	2	196,905,152	196,923,530	18.38	0.032	0.571	26	895	45	884
S1370	12	87,203,044	87,313,656	110.61	0.033	0.744	113	808	147	782
S26017	12	43,871,077	43,932,533	61.46	0.034	3.731	11	910	3	926
S20366	9	127,896,163	127,926,763	30.60	0.034	2.893	17	904	6	923
S25279	11	107,375,709	107,469,157	93.45	0.034	0.250	3	918	12	917
S6910	2	167,328,609	167,449,529	120.92	0.036	1.622	55	866	35	894
S5205	8	112,894,399	112,951,620	57.22	0.037	0.650	45	876	68	861
S1416	12	86,947,028	86,972,486	25.46	0.037	0.747	111	810	144	785
S12894	2	200,003,795	200,076,595	72.80	0.037	0.496	15	906	30	899
S28335	13	48,347,645	48,391,086	43.44	0.037	4.574	9	912	2	927
S28496	9	26,323,942	26,623,841	299.90	0.037	4.574	9	912	2	927
S28833	2	202,028,297	202,133,831	105.53	0.037	4.574	9	912	2	927
S28834	2	201,511,628	201,564,833	53.21	0.037	4.574	9	912	2	927
S28965	1	115,587,687	115,592,174	4.49	0.037	4.574	9	912	2	927
S20450	6	96,300,666	96,352,261	51.60	0.038	0.352	6	915	17	912
S31580	3	115,042,710	115,069,466	26.76	0.038	7.107	7	914	1	928
S27423	19	19,725,309	19,737,112	11.80	0.038	0.200	2	919	10	919
S10644	10	48,262,744	48,270,202	7.46	0.039	1.849	36	885	20	909
S30018	14	51,978,575	52,903,676	925.10	0.039	0.125	1	920	8	921
S1364	12	87,161,152	87,193,257	32.11	0.039	0.751	114	807	147	782
S22587	11	64,821,266	64,833,494	12.23	0.040	2.852	14	907	5	924
S5670	8	113,184,531	113,258,104	73.57	0.040	0.646	42	879	64	865
S17359	4	171,033,461	171,040,691	7.23	0.040	2.385	21	900	9	920
S12429	2	200,173,497	200,247,879	74.38	0.041	0.512	16	905	31	898
S19551	4	60,080,618	60,143,671	63.05	0.042	2.626	18	903	7	922
S11606	2	168,265,559	168,286,668	21.11	0.043	1.881	33	888	18	911
S5403	8	113,591,733	113,613,045	21.31	0.044	0.656	44	877	66	863
S13135	2	98,841,055	98,881,076	40.02	0.044	1.981	29	892	15	914

S13584	3	165,780,254	165,818,204	37.95	0.045	0.497	14	907	28	901
S13608	2	200,394,844	200,422,724	27.88	0.045	0.497	14	907	28	901
S8979	2	131,508,017	131,521,911	13.89	0.046	1.701	43	878	26	903
S11963	4	9,596,616	9,780,939	184.32	0.046	0.527	17	904	32	897
S11970	3	165,443,716	165,464,246	20.53	0.046	0.527	17	904	32	897
S2912	8	112,557,997	112,567,357	9.36	0.047	0.711	69	852	95	834
S24474	14	64,999,089	65,307,473	308.38	0.047	3.053	12	909	4	925
S24501	13	84,397,683	84,403,995	6.31	0.047	3.053	12	909	4	925
S11184	5	50,599,580	50,631,480	31.90	0.047	1.836	34	887	19	910
S1721	8	111,915,287	112,012,520	97.23	0.048	0.747	99	822	129	800
S8382	10	69,581,580	69,587,248	5.67	0.048	0.602	28	893	46	883
S23688	19	43,450,782	43,456,912	6.13	0.048	0.307	4	917	13	916
S12484	16	67,518,483	67,563,851	45.37	0.049	1.921	30	891	16	913

I then repeated the analysis of common regions of ROHs using the low-LD panel, which resulted in 99 overlapping regions. However, none of the regions detected were significantly associated with disease (all P values were greater than 0.12, see Table 6.8), although, just 51 individuals were detected with ROHs using this SNP panel.

Table 6.8 The recurrent ROH regions detected in the low-LD panel

The common ROH regions that identified in the low-LD panel with a consensus region spanning more than 1 SNP, to ensure a reasonable overlap, are listed here. The start and end positions given provide the size and location of the consensus region. P values were calculated by Fisher's exact test, as cell values were less than 5.

ROH	Chr	Start position (bp)	End position (bp)	Size (kb)	P value	OR	ROH aff	No ROH aff	ROH ctrl	No ROH ctrl
S59	1	222,319,421	225,414,014	3094.59	0.123	NA	3	918	0	929
S56	2	14,717,412	18,608,193	3890.78	0.123	NA	3	918	0	929
S50	3	6,123,891	6,200,539	76.648	0.123	NA	3	918	0	929
S17	20	15,644,084	19,881,188	4237.1	0.123	NA	3	918	0	929
S2	6	160,435,702	163,636,747	3201.05	0.216	4.048	4	917	1	928
S189	1	30,250,660	36,502,173	6251.51	0.248	NA	2	919	0	929
S179	1	205,014,126	207,743,233	2729.11	0.248	NA	2	919	0	929
S172	2	20,272,631	20,665,545	392.914	0.248	NA	2	919	0	929
S169	2	53,773,212	64,610,788	10837.6	0.248	NA	2	919	0	929
S168	2	66,726,926	74,409,108	7682.18	0.248	NA	2	919	0	929
S165	2	238,405,149	242,759,899	4354.75	0.248	NA	2	919	0	929
S158	3	109,072,371	115,890,321	6817.95	0.248	NA	2	919	0	929
S153	4	12,957,031	17,095,582	4138.55	0.248	NA	2	919	0	929
S148	5	5,878,953	9,546,723	3667.77	0.248	NA	2	919	0	929
S147	5	32,029,579	35,151,101	3121.52	0.248	NA	2	919	0	929
S119	9	79,101,054	88,121,083	9020.03	0.248	NA	2	919	0	929
S118	9	88,328,421	88,865,339	536.918	0.248	NA	2	919	0	929

S117	9	89,871,652	98,673,380	8801.73	0.248	NA	2	919	0	929
S116	9	98,805,064	105,474,750	6669.69	0.248	NA	2	919	0	929
S105	13	30,800,042	30,860,283	60.241	0.248	NA	2	919	0	929
S104	13	32,451,508	37,244,250	4792.74	0.248	NA	2	919	0	929
S102	13	67,233,750	72,717,898	5484.15	0.248	NA	2	919	0	929
S101	13	87,078,943	91,920,806	4841.86	0.248	NA	2	919	0	929
S84	16	64,481,870	72,458,458	7976.59	0.248	NA	2	919	0	929
S76	19	57,705,611	58,318,325	612.714	0.248	NA	2	919	0	929
S75	19	61,764,245	63,740,123	1975.88	0.248	NA	2	919	0	929
S74	20	834,856	1,707,590	872.734	0.248	NA	2	919	0	929
S70	21	21,405,207	23,130,703	1725.5	0.248	NA	2	919	0	929
S65	22	28,166,196	32,649,309	4483.11	0.248	NA	2	919	0	929
S34	8	59,228,801	62,111,121	2882.32	0.250	NA	0	921	3	926
S33	8	79,197,411	79,998,890	801.479	0.250	NA	0	921	3	926
S32	8	82,365,772	85,293,798	2928.03	0.250	NA	0	921	3	926
S31	8	90,706,741	95,772,145	5065.4	0.250	NA	0	921	3	926
S23	15	90,370,621	92,417,738	2047.12	0.250	NA	0	921	3	926
S22	15	94,476,582	98,418,122	3941.54	0.250	NA	0	921	3	926
S4	16	79,228,642	79,765,704	537.062	0.372	3.033	3	918	1	928
S166	2	234,278,945	235,754,217	1475.27	0.500	NA	0	921	2	927
S157	3	174,635,461	181,850,955	7215.49	0.500	NA	0	921	2	927
S152	4	57,767,664	71,607,301	13839.6	0.500	NA	0	921	2	927
S145	5	114,497,623	121,663,274	7165.65	0.500	NA	0	921	2	927
S142	6	23,512,255	37,804,574	14292.3	0.500	NA	0	921	2	927
S140	6	138,435,007	147,061,448	8626.44	0.500	NA	0	921	2	927
S134	7	72,837,254	80,476,384	7639.13	0.500	NA	0	921	2	927
S112	10	104,132,284	109,730,376	5598.09	0.500	NA	0	921	2	927
S110	11	11,549,487	16,340,006	4790.52	0.500	NA	0	921	2	927
S97	14	79,393,165	86,565,126	7171.96	0.500	NA	0	921	2	927
S69	21	31,193,930	33,352,040	2158.11	0.500	NA	0	921	2	927
S68	21	41,813,926	42,158,582	344.656	0.500	NA	0	921	2	927
S60	1	148,732,419	153,426,590	4694.17	0.623	2.020	2	919	1	928
S58	1	226,518,133	228,465,414	1947.28	0.623	2.020	2	919	1	928
S57	1	233,969,275	236,996,468	3027.19	0.623	2.020	2	919	1	928
S49	3	54,961,511	59,878,234	4916.72	0.623	2.020	2	919	1	928
S48	3	60,102,656	60,423,030	320.374	0.623	2.020	2	919	1	928
S47	4	7,704,995	8,272,605	567.61	0.623	2.020	2	919	1	928
S28	12	1,532,313	2,078,651	546.338	0.623	2.020	2	919	1	928
S27	12	4,077,029	4,854,862	777.833	0.623	2.020	2	919	1	928
S24	15	66,884,660	68,265,584	1380.92	0.623	2.020	2	919	1	928
S15	2	227,389,530	227,928,257	538.727	0.625	0.336	1	920	3	926
S8	8	3,567,552	6,118,677	2551.12	0.625	0.336	1	920	3	926
S63	1	39,259,917	44,540,668	5280.75	1.000	0.504	1	920	2	927
S186	1	46,211,546	53,291,423	7079.88	1.000	1.009	1	920	1	928
S185	1	59,018,017	61,956,300	2938.28	1.000	1.009	1	920	1	928
S62	1	110,363,064	110,801,447	438.383	1.000	0.504	1	920	2	927
S55	2	37,933,149	38,792,757	859.608	1.000	0.504	1	920	2	927
S16	2	217,848,519	219,077,713	1229.19	1.000	1.009	2	919	2	927
S159	3	73,508,536	76,167,310	2658.77	1.000	1.009	1	920	1	928
S156	4	2,223,947	5,260,284	3036.34	1.000	1.009	1	920	1	928
S151	4	78,087,439	89,560,376	11472.9	1.000	1.009	1	920	1	928
S150	4	100,687,487	107,856,853	7169.37	1.000	1.009	1	920	1	928

S149	4	187,605,886	188,097,628	491.742	1.000	1.009	1	920	1	928
S146	5	39,516,957	55,253,515	15736.6	1.000	1.009	1	920	1	928
S12	6	47,374,655	51,251,512	3876.86	1.000	1.009	2	919	2	927
S42	6	151,339,981	152,570,934	1230.95	1.000	0.504	1	920	2	927
S10	6	165,855,198	166,156,688	301.49	1.000	1.009	2	919	2	927
S135	7	16,425,763	17,265,076	839.313	1.000	1.009	1	920	1	928
S133	7	86,773,516	95,199,567	8426.05	1.000	1.009	1	920	1	928
S9	8	2,761,464	3,340,467	579.003	1.000	1.009	2	919	2	927
S35	8	27,460,271	28,212,815	752.544	1.000	0.504	1	920	2	927
S121	8	134,915,914	135,861,295	945.381	1.000	1.009	1	920	1	928
S120	9	66,560,586	74,462,048	7901.46	1.000	1.009	1	920	1	928
S115	10	24,435,153	24,448,404	13.251	1.000	1.009	1	920	1	928
S1	10	66,885,229	68,293,538	1408.31	1.000	0.672	2	919	3	926
S111	10	113,934,403	118,083,598	4149.19	1.000	1.009	1	920	1	928
S106	12	127,587,138	129,161,657	1574.52	1.000	1.009	1	920	1	928
S103	13	38,251,238	42,624,100	4372.86	1.000	1.009	1	920	1	928
S100	14	22,378,826	22,575,950	197.124	1.000	1.009	1	920	1	928
S99	14	23,109,399	28,926,702	5817.3	1.000	1.009	1	920	1	928
S98	14	31,265,290	33,914,127	2648.84	1.000	1.009	1	920	1	928
S26	15	52,896,124	55,431,748	2535.62	1.000	0.504	1	920	2	927
S25	15	58,996,303	60,143,678	1147.38	1.000	0.504	1	920	2	927
S87	16	7,801,478	8,865,964	1064.49	1.000	1.009	1	920	1	928
S81	17	52,877,449	59,397,940	6520.49	1.000	1.009	1	920	1	928
S80	17	67,044,820	67,672,674	627.854	1.000	1.009	1	920	1	928
S79	17	67,904,675	68,971,738	1067.06	1.000	1.009	1	920	1	928
S78	18	70,952,589	73,600,325	2647.74	1.000	1.009	1	920	1	928
S77	19	55,897,865	56,241,010	343.145	1.000	1.009	1	920	1	928
S71	21	15,237,709	20,556,909	5319.2	1.000	1.009	1	920	1	928
S67	21	42,206,074	46,909,417	4703.34	1.000	1.009	1	920	1	928
S66	22	25,525,496	25,720,163	194.667	1.000	1.009	1	920	1	928
S64	22	35,274,800	36,131,803	857.003	1.000	1.009	1	920	1	928

6.3.2.4 Comparison of overlapping regions with detected copy number variation (CNV)

In order to better classify the nature of these homozygous regions and determine whether they were regions of autozygosity or actually hemizygous CNVs, I compared the positions of the common overlapping ROHs with CNVs detected (by my colleague Jean-Baptiste Cazier) in the EngP1 dataset using QuantiSNP (Colella *et al.* 2007). I searched for CNVs that had at least a 90% overlap with the detected common ROH regions. However, none of the ROHs, using either the 500K or the low-LD SNP panel, could be explained by a CNV.

6.4 Discussion

The recent studies by Assié *et al.* and Bacolod *et al.* (Assie *et al.* 2008; Bacolod *et al.* 2008) reported evidence to suggest that an increase in homozygosity/autozygosity in cases is associated with an increased risk of CRC and other cancers. The results have been interpreted as consistent with the results of some of the studies on inbred populations that highlighted the detrimental effects increased homozygosity has on health, including increased cancer risk. The results led Bacolod *et al.* to form a new homozygosity based model for the progression of CRC (Bacolod *et al.* 2009). However, although the neatness of the explanation that increased homozygosity causes increased cancer risk is attractive, there are several criticisms that can be made of the design of these studies. The observation that increased homozygosity increases cancer risk was largely discovered by studying a small number of individuals from an isolated inbred population or one large family showing a high level of consanguinity. Therefore, it is not clear whether extrapolating these observations to a (largely outbred) population is relevant in terms of cancer risk, especially as the effect of inbreeding may be confounded by founder effects. If the founder has an increased risk through some rare deleterious mutation and then subsequent generations show inbreeding, the mutation is more likely to be inherited in a homozygous form and lead to an increase of the disease in the family. However, this is dependent on the founder. Therefore, it does not follow that an increase in homozygosity through inbreeding will lead to an increase in disease risk in the population. Additionally, with the recent studies based on genetically determined levels of homozygosity (Bacolod *et al.* 2008), the study sample sizes were small (74 cases and 146 plus 118 controls) and cases and controls were heterogeneous or ethnically unmatched. This study included Ashkenazi Jewish cases, which one would expect to have increased levels of homozygosity, and compared them with non-Jewish controls. Another issue to be considered is the effect of LD on the test statistics, which may have inflated any differences between cases and controls. For example, there are 58,136 SNPs on the Affymetrix 50K *Xba* I SNP array used in the analysis by Bacolod *et al.*, but only 11,142 SNPs remain once those in high pairwise LD (r^2>0.1) are filtered out.

The ideal method to analyse level of autozygosity in relation to CRC risk is not clear, especially as most western populations are not particularly inbred and therefore the number of generations between common ancestors is likely to be large and the levels of detectable autozygosity very small. Additionally, ROHs can arise through events other than autozygosity. Equally, controlling for covariates in individuals with detectable autozygosity is extremely difficult.

In an attempt to overcome some of these issues and test the findings of the previous molecular studies, I have performed an analysis on overall levels of homozygosity and runs of consecutive homozygous markers using a relatively large dataset of ethnically matched cases and controls that have been genotyped for a dense genome-wide panel of SNPs. These samples were utilised as the discovery phase for our GWAS for CRC (see Chapter 3), robustly identifying several predisposition SNPs, and show no evidence for population stratification. In this study, both including and excluding SNPs in pairwise

LD, any evidence to suggest that an increase in homozygosity is associated with CRC risk was very limited. I did not find the cases to be significantly more homozygous, or inbred, than controls. Additionally, I did not find cases to have a significantly increased number of ROHs or total ROH size using several ROH calling criteria.

As discussed in Chapter 4, the use of the imputed data in VQ58 for the association by homozygosity meta-analysis was sub-optimal, as the cases were imputed and the controls were genotyped for SNPs missing in VQ58. However, this did not appear to cause problems in any of the 10 imputed SNPs analysed in this study.

From the results of these analyses, I can only conclude that, in the UK population where inbreeding levels are low, there is no compelling evidence to support the hypothesis that increased levels of homozygosity are associated with an increase in CRC risk. The analysis did not distinguish between the causes of stretches of homozygosity, such as uniparental isodisomy, autozygosity or hemizygosity. Equally, the results did not identify any recessive acting alleles, although the effect of such alleles in inbred groups is not discounted. The Illumina Hap550 SNP array was not designed to tag variations with a frequency of less than 5% and so low frequency recessive alleles may go undetected. It could also be that deleterious recessive alleles are too heterogeneous or too rare in the population to be identified by autozygous regions in most white European populations. Other studies have been performed to assess the association of increased homozygosity with cancer risk with similar findings. Hosking and colleagues performed a comparable analysis, using genome-wide SNP data, of the effect of ROHs on the risk of childhood Acute Lymphoblastic Leukaemia

(ALL)(Hosking *et al.* 2010). Using 228,714 SNPs in approximate linkage equilibrium and 3,180 samples, of which 824 were cases, the researchers found no evidence of an association between total size of ROH and increased ALL risk. The same group also studied the effect of homozygosity in breast and prostate cancer, where there was also no strong evidence of an association with disease risk (Enciso-Mora *et al.* 2010). Together these results indicate that increased levels of homozygosity, whatever their cause, are unlikely to cause significantly increased cancer risk in outbred populations.

Chapter 7. The detection of moderate penetrance or rare susceptibility alleles

7.1 Introduction

This chapter is separated into three main sections that describe some additional methods to identify CRC susceptibility loci. The first section covers a revised linkage analysis of several large families from which some of the affected individuals from EngP1 belong. In the second section, I describe an analysis of the somatic chromosomal aberrations in the tumours of the affected individuals from the linkage families and a comparison of the identified linkage regions with those where loss of heterozygosity (LOH) was detected. The final section describes an analysis of segments shared identical by descent (IBD) in the group of Jewish individuals identified in the PCA analysis in Chapter 2 and in the ScotP1 dataset, using the genotypes from the GWA study.

7.2 Linkage analysis for the detection of CRC susceptibility loci

The EngP1 samples belong to a collection of CRC families that were recruited as part of the Colorectal Tumour Gene Identification Study (CORGI). The families recruited to this study were required to meet affection status criteria (described fully in the Materials and Methods) to include a minimum of three members confirmed to be affected, the exclusion of mutations known to cause Mendelian CRC syndromes, such as APC and HNPCC, and be affected with either significant adenomas or CRC before the age of 70 years.

7.2.1 Summary of previous linkage analysis results on CORGI families

As a result of the CORGI study, a number of large families (n=69) with up to 10 affected members in each family that demonstrated evidence of a mostly dominant inheritance, were identified that were suitable for linkage analysis. Available individuals from these families were genotyped using the Affymetrix 10K genome wide SNP arrays and analysed in a combined linkage study. The results identified significant linkage peaks on chromosome 3q21-24 and chromosome 18q21 (Kemp *et al.* 2006). The region of linkage at 3q21-24 was later replicated by an independent study (Picelli *et al.* 2008).

Most recently 34 additional families were included in the analysis, which refined the chromosome 3 linkage region to 3q22 and found suggestive evidence of linkage to 18q21 (Papaemmanuil *et al.* 2008). In this combined analysis, although two significant linkage peaks were discovered, mutation screening of candidate genes within these regions failed to identify any causal mutations to explain the signal.

7.2.2 Reasons for a reviewed analysis of these families

The results of the combined linkage analysis showed evidence of heterogeneity at the disease locus between families included in the study. This could result in loss of a linkage signal for rare loci specific to individual families. In addition, a number of families within this dataset show evidence of inheritance patterns with moderate penetrance in the pedigree structure and a phenotype more similar to known Mendelian predisposition conditions (The family pedigrees are given in the Appendix figure 9.4). For example, one of the most intriguing of the CORGI linkage families is family 336, where the affected members present with a multiple adenoma phenotype more similar to AFAP, with a young age of diagnosis and several cases of endometrial cancer, which can be indicative of HNPCC. However, there was no evidence of microsatellite instability (often associated with mutations in mismatch repair genes) in the tumours of affected individuals in this family. To date no mutations in known Mendelian cancer predisposition genes have been detected to explain the variance in CRC risk in this family.

Therefore, to attempt to identify moderate penetrance disease loci specific to certain families, I performed a separate linkage analysis on eight of the individual families. These families have a relatively small number of generations and thus have limited power and resolution to detect a disease locus. However, most of the affected individuals also presented with a number of adenomas, which were available for analysis (see table 7.1 for summary pathology data for each family). It is likely that genes that increase the risk of adenomas are also involved in CRC risk. These adenoma samples can be classed as affected individuals and their analysis will provide increased power to detect disease loci.

The adenomas were included in a loss of heterozygosity analysis (LOH), which was used to compare any detected somatic alterations, which may point to the presence of tumour suppressor genes, with the peaks identified in the linkage analyses. The rationale behind this analysis is based on several published observations in tumours that are described below.

7.2.3 Loss of Heterozygosity (LOH)

As cancer cells are the result of an accumulation of mutations and chromosomal abnormalities that have been acquired somatically by the cell, the presence of genetic alterations in tumour cells can be used to highlight disease loci. LOH is the process whereby a cell, heterozygous for an inherited recessive mutation inactivating one allele, somatically acquires another mutation that inactivates the second allele and renders the cell homozygous for the disease locus.

This process was observed in a study of the childhood cancer retinoblastoma, which is caused by a mutation affecting the tumour suppressor gene, Rb1. It was noted that inherited cases were often bilateral (affected both eyes), whereas sporadic cases where unilateral. The observation led Alfred Knudson to suggest the two-hit hypothesis where the development of a tumour required two mutational events, which could either both occur in the somatic cells (in the sporadic case) or one could be inherited in the germline and the other acquired later (Knudson 1971). The hypothesis was further investigated by Cavenee and colleages who showed that retinoblastoma could be the result of homozygosity for the mutant allele at the disease locus (Cavenee *et al.* 1983). The study provided evidence that chromosomal events in somatic cells, such as LOH, could lead to the formation of tumours through the expression (or inactivation) of recessive alleles. LOH can occur by deletion, which causes a change in copy number, but one of the most common methods is mitotic recombination, which is a means of repairing double-stranded breaks in DNA (Valerie and Povirk 2003). Mitotic recombination can result in a reduction to homozygosity in the somatic cell, but not a reduction in copy number. A good example of this process in CRC is seen in FAP, where the tumour suppressor gene, *APC,* undergoes two mutational hits (one inherited in the germline) to inactivate both copies of the gene. Copy neutral LOH at the *APC* locus is seen in 85% of sporadic CRC cases (Howarth *et al.* 2009).

LOH is frequently reported in colorectal tumour samples and can be used to map the location of tumour suppressor (TS) genes. These genes function to suppress cellular processes that might lead to the development of a cancerous cell, for example by inhibiting the cell cycle or inducing apoptosis. However, loss of function mutations in both alleles of TS genes removes this control, allowing tumours to develop unchecked. Therefore, an inherited mutation in a TS gene will give the carrier a predisposition to developing cancer.

SNP LOH has been performed on colorectal cancers previously (Gaasenbeek *et al.* 2006; Howarth *et al.* 2009) and has successfully identified LOH at many of the sites common to the colorectal tumorigenesis model (Fearon and Vogelstein 1990). These sites include chromosome 18q, which is frequently seen in late adenomas and cancers, 5q and 17p, which is more common in cancers.

Table 7.1 Summary of pathology data for individuals in each family

The data shown includes all affected individuals in each of the families included in the linkage and LOH analysis and shown in the pedigrees in figure 7.1 below. HPP stands for hyperplastic polyp. The number of adenomas that had the more dysplastic morphology of a TVA (tubulovillous adenoma) is indicated in brackets. This gives an indication of the differences in phenotype and severity of phenotype between the families.

Family	Individual	Diagnosis	Age of diagnosis	No of Adenomas	No of HPP
323	0120_301	Adenoma	41	4 (1 TVA)	2
323	0120_304	Adenoma	44	11	1
323	0120_305	Adenoma	56	1	-
323	0120_311	Adenoma	60	4	-
323	0120_401	Adenoma	34	2	-
323	0120_303	Hyperplastic polyp	50	-	1
323	0120_202	Adenoma	64	1	1
323	0120_201	CRC	71	1 (TVA)	-
336	0122_301	Adenoma	43	9 (3 TVA)	-
336	0122_303	Adenoma Endometrial Cancer	54 45	15	-
336	0122_304	Adenoma, BCC Endometrial Cancer	54 52	35	5

336	0122_403	Adenoma	29	1	-
336	0122_404	Adenoma	30	2	10
336	0122_405	CRC	28	24 (1 TVA)	3
336	0122_402	Adenoma Brain tumour	34 26	3	2
282	0088_522	Adenoma	35	4	-
282	0088_401	CRC	62	4	3
282	0088_501	Adenoma (unconfirmed)		-	-
282	0088_407	Adenoma	73	1	1
282	0088_408	Adenoma	61	3	-
282	0088_409	Two CRCs	60/64	1	-
282	0088_402	CRC	81	-	-
294	0065_301	Adenoma	49	2 TVA	2 (1 Serrated)
294	0065_302	Adenoma	57	3 (1 TVA)	-
294	0065_304	CRC	51	1 TVA	-
294	0065_311	Adenoma	36	1	-
294	0065_307	Hyperplastic polyps	53	-	9
294	0065_314	Hyperplastic polyps	35	-	1
294	0065_309	Adenoma	50	3	32 (7 Serrated)
294	0065_308	Adenoma	46	1 TVA	1
294	0065_312	Hyperplastic polyps	37	-	2
450	0162_103	Adenoma	58	2 TVA	-
450	0162_201	Adenoma	31	1	1
450	0162_203	Adenoma	?	1	-
450	0162_204	Adenoma	?	1	-
450	0162_301	Adenoma	14	1	-
450	0162_202	Adenoma	29	1 TVA	-
450	0162_104	CRC	52	1 (unconfirmed)	-
329	0109_302	Adenoma	45	6 (1 TVA)	2
329	0109_301	CRC	50	7 (2 TVA)	Multiple
329	0109_303	CRC	44	-	-
329	0109_304	CRC	54	-	-
329	0109_401	Adenoma	27	2 (unconfirmed)	-
329	0109_201	CRC	51	-	-
377	0125_102	CRC (unconfirmed)	47	-	-
377	0125_201	Adenoma	35	2 (1 TVA)	-
377	0125_202	Adenoma	46	2	-
377	0125_204	CRC	40	13 (4 TVA)	1
377	0125_205	Adenoma	38	8	1
377	0125_302	Adenoma	16	6	-
377	0125_301	Adenoma	27	1	-
326	0114_304	CRC (unconfirmed)	27	-	-
326	0114_301	Adenoma	58	1 TVA	-
326	0114_302	Adenoma	37	1	14
326	0114_303	Adenoma	52	3	4
326	0114_401	Adenoma	27	1	3
346	0127_202	CRC	61	-	-
346	0127_204	CRC	57	-	-
346	0127_203	CRC	44	-	-
346	0127_205	CRC	33	-	-
346	0127_206	CRC	64	-	-
346	0127_301	CRC	29	-	-
346	0127_305	CRC	49	-	-

346	0127_304	CRC	36	-	-
346	0127_401	Adenoma	20	1	9 (1 Serrated)
346	0127_402	Adenoma	21	-	1

7.2.4 Strategy for the single family linkage analysis

I performed a single family linkage analysis on the most potentially informative and interesting eight families from the published analyses. These families were all genotyped using the Affymetrix Genechip Mapping 10K linkage array. All, but two of these families (326 and 346) also had tumours available for inclusion in the subsequent LOH study.

The analysis was undertaken using both multipoint parametric (dominant and recessive) and non-parametric models to maximise the chance of identifying any regions of linkage. However, the mode of inheritance for these families is more akin to a dominant model. Full details of the methods and parameters used for the analysis are given in the Materials and Methods Chapter. In order to gauge the maximum LOD scores to expect in the presence of linkage for each family, I calculated estimated maximum LOD scores by performing 1000 iterations of simulating genotypes and running the linkage analysis using the same parameters as described (see Table 7.2).

Table 7.2 The estimated maximum LOD score for individual families

The LOD scores given in the table were the maximum LOD scores obtained under a dominant and recessive model for the each family after performing linkage analysis on simulated genotypes. 1000 sets of simulated genotypes were generated for each family based on the pedigree structure, allele frequencies, and assuming a disease associated locus at 50.53cM. The maximum LOD score obtained was taken as an estimate for the highest expected LOD score in each family.

Family ID	Max LOD Dominant Model	Max LOD Recessive Model
323	1.770	1.110
336	1.739	1.695
329	0.758	1.628
282	0.661	1.459
377	0.777	1.412
346	1.038	0.616

294	1.231	2.647
326	0.684	1.277

7.2.5 Linkage analysis results

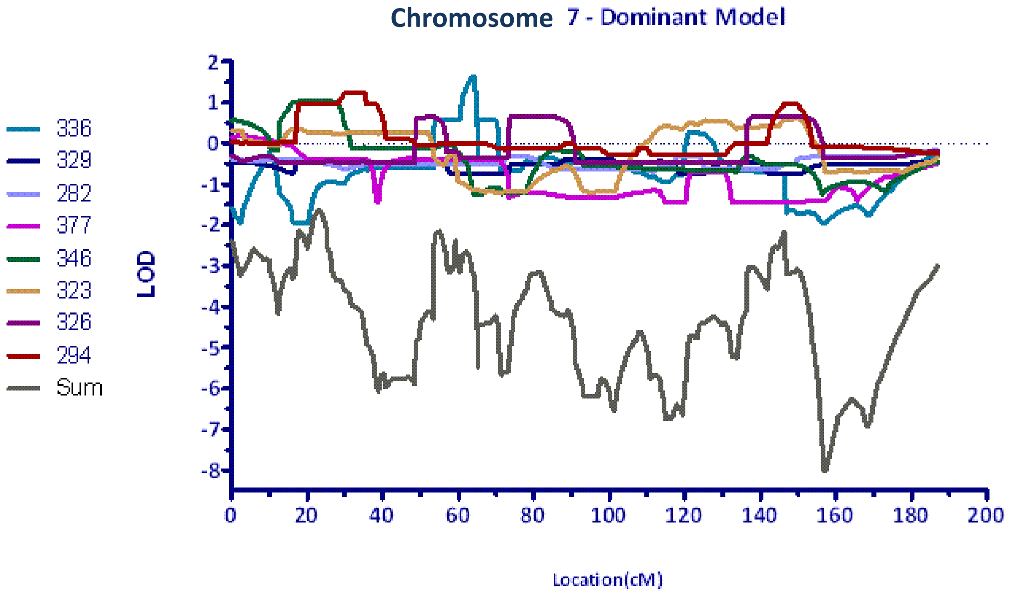
The non-parametric multipoint analysis did not identify any regions of linkage with LOD scores greater than 1 in any family and, therefore, the results detailed below only cover the results of the parametric analyses. This is not entirely unexpected as the non-parametric model suffers a loss in power in the absence of a pre-defined genetic model.

The single family analyses for the recessive and dominant models showed little overlap between families, indicating heterogeneity at the disease loci. However, regions were identified in individual families with LOD scores suggestive of linkage, which warrant further study. In the interests of space, only the chromosomes showing the highest LOD scores (greater than 1.4) are shown here in Figure 7.1 and Figure 7.2 (and the actual regions given in Table 7.7, in Section 7.3.3.3). However, the results for the remaining chromosomes under both models are given in the appendix.

Although the LOD scores are low in relation to the accepted level of significance for linkage (LOD>3), the peaks that were identified were close to the maximum LOD score obtained for each family in the analyses using simulated genotypes. With the exception of the region on chromosome 3 at 41.9-46.8cM in family 336 (where the haplotype was absent from 0122_405), the risk haplotypes responsible for each of the peaks with LOD>1.4 segregated perfectly with disease and were present in each affected member and absent from each unaffected member genotyped.

Figure 7.1 Dominant model linkage results with LOD>1.4

The plots included in this figure, for chr7 and 10, detail the regions of linkage detected with LOD scores greater than 1.4, in any of the families analysed. The LOD scores are based on an individual family analysis using a parametric dominant model and highlight the heterogeneity between the different families. The complete results are given in the appendix.



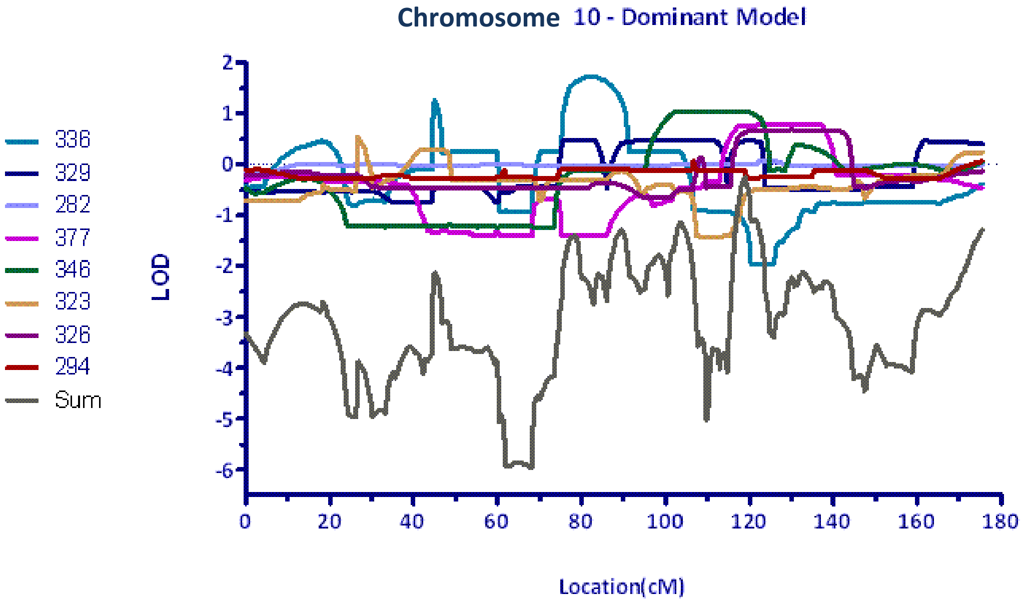
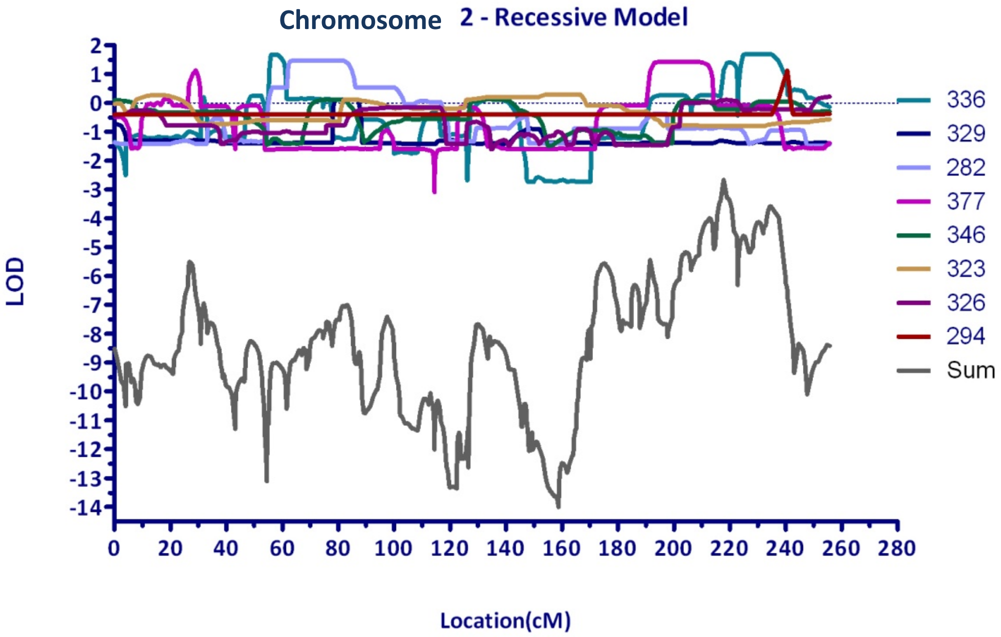
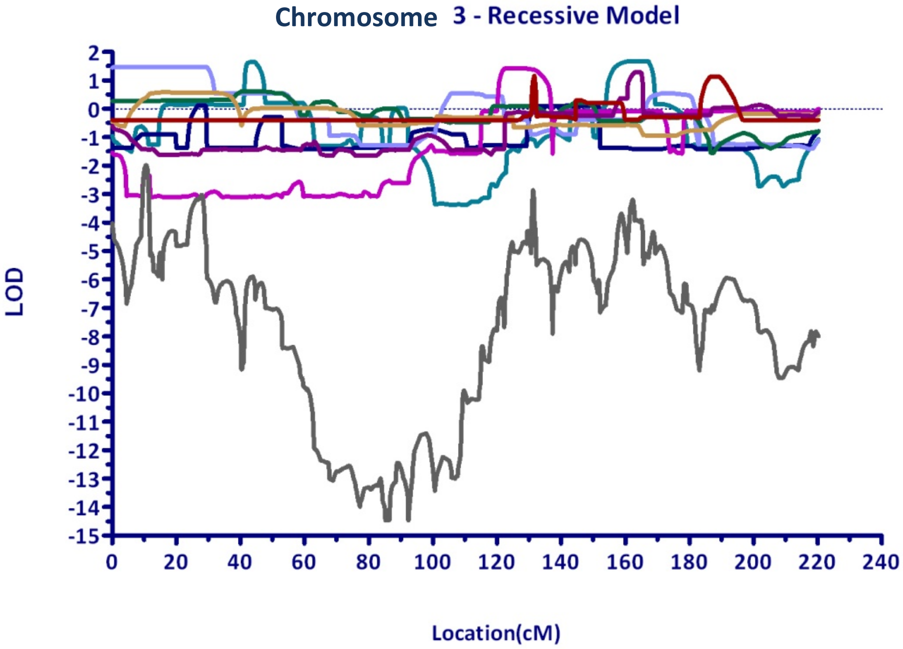
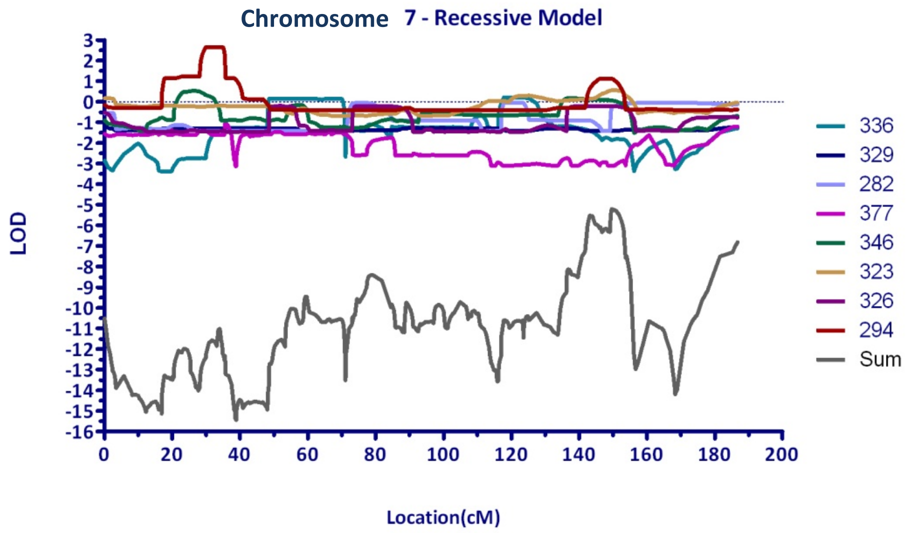


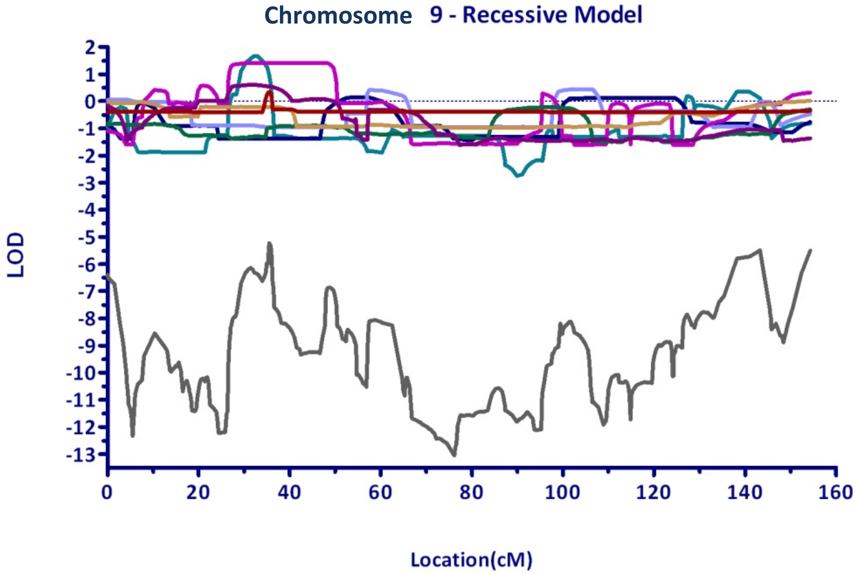
Figure 7.2 Recessive Model Linkage results with LOD>1.4

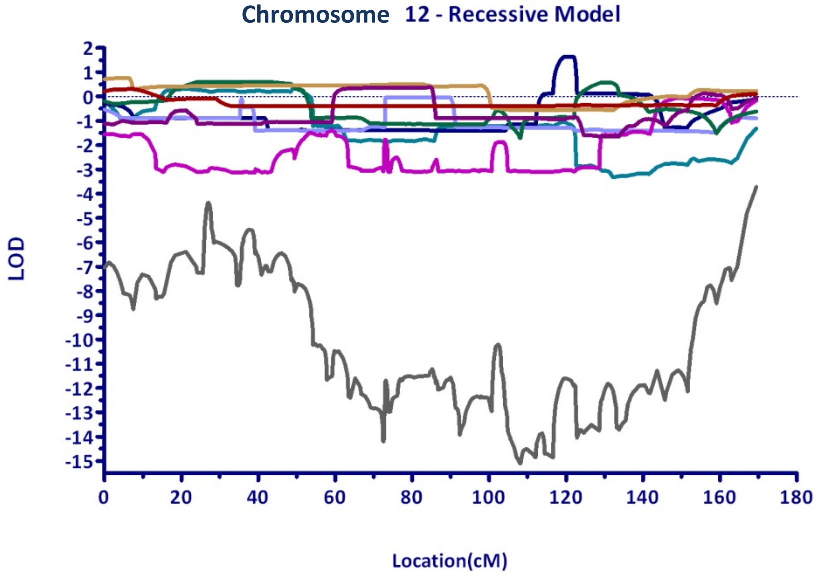
The plots included in this figure detail the regions of linkage detected with LOD scores greater than 1.4, in any of the families analysed. The LOD scores are based on an individual family analysis using a parametric recessive model and highlight the heterogeneity between the different families. The LOD scores from each family were added together to produce the grey line to give the LOD score obtained if all families were combined. The complete results are given in the Appendix.






Chromosome 7 - Recessive Model
LOD
3
2
1
0
-1
-2
-3
-4
-5
-6
-7
-8
-9
-10
-11
-12
-13
-14
-15
-16
0
20
40
60
80
100
120
140
160
180
200
Location(cM)
336
329
282
377
346
323
326
294
Sum


Chromosome 9 - Recessive Model
LOD
2
1
0
-1
-2
-3
-4
-5
-6
-7
-8
-9
-10
-11
-12
-13
0
20
40
60
80
100
120
140
160
Location(cM)


Chromosome 12 - Recessive Model
LOD
2
1
0
-1
-2
-3
-4
-5
-6
-7
-8
-9
-10
-11
-12
-13
-14
-15
0
20
40
60
80
100
120
140
160
180
Location(cM)

7.3 LOH analysis of linkage family tumours

I performed a genome-wide loss of heterozygosity (LOH) analysis, using SNP LOH, on the tumours of individuals from families included in the single family linkage analysis to search for somatic changes that may indicate the presence of novel tumour suppressor genes that could contribute to disease susceptibility. The results of this analysis were compared to the loci detected in the linkage analysis above to determine if they coincided with regions of LOH. In this way, LOH can be used to fine map a linkage region if it can be shown that a region of LOH segregates with disease.

7.3.1 Study design

Families where more than one tumour sample was available in multiple family members were further analysed in a SNP LOH experiment (see Table 7.3). Owing to the interesting phenotype of affected individuals, an additional family (family 450) was added to the LOH analysis that was not included in the linkage analysis.

The presence of microsatellite instability (MSI) in the tumours of each family was assessed by screening BAT25 and BAT26 in the two most dysplastic high grade adenomas or cancers from different individuals within each family. MSI is associated with mismatch repair gene mutations, which can indicate HNPCC. None of the families included in this study demonstrated MSI in the tumours tested.

I attempted to limit the included samples to adenomas or tubulovillous adenomas, rather than cancer samples. This is because cancers generally undergo many additional changes through mutations accumulated during clonal expansion and CIN, which can lead to aneuploidy, which would produce ambiguous results (two cancer samples were included). Tumours that were large enough to extract a sufficient quantity of DNA were analysed, 1µg at 50ng/µl of DNA was required for optimal results. However, a number of the DNA samples from smaller adenomas were only 25-30ng/µl (the majority of these samples were successful). The method for the extraction of DNA from paraffin embedded blocks is given in the materials and methods chapter. The tumour DNA genotyped for 6,056 SNPs using the Illumina GoldenGate Human linkage panel following the manufacturer's protocol.

Table 7.3 Summary of the tumours analysed from each family included in the LOH analysis

I was not able to analyse the tumours of all the families included in the single family linkage analysis and this table contains the families that were included in the LOH analysis and the number of tumours available and the number of different individuals from which tumours were studied. None of these families showed evidence of microsatellite unstable tumours.

Linkage family ID	No. of tumours	Number of individuals analysed
336	31	6
323	9	5
377	29	6
450	7	3
282	5	2
294	12	5
329	2	1

7.3.2 Initial LOH study results

An initial scan, by eye, of the plots of B allele frequency and log R ratio produced in the Illumina Genomestudio package from the raw intensity data, identified the presence of a number of large regions of LOH covering the whole or part of a chromosome arm (see Table 7.4, which also details the pathology of the tumour samples that were analysed for each sample). The most common regions of LOH were located on chromosome 5q (n=6), 18q (n=7) and 19q (n=9). There were three tumours from the same individual in family 294 with loss of 5q, although this was not detected in any other member of the family. There were seven tumours from families 336, 377 and 329, which showed loss of 18q, which is commonly identified in colorectal tumours.

Table 7.4 Tumours Analysed from each family and regions of LOH identified from the B allele frequency and log R ratio information

This table details the individuals from each of the linkage families that were included in the analysis, the tumours analysed and the regions of LOH detected in the tumour samples. A "-" in the LOH or CNV (gain) column indicates that no obvious changes were identified, otherwise the general chromosomal region is given. The 'type' column gives the pathology of the sample, where TA is tubular adenoma, TVA is tubulovillous adenoma, atypical TVA is a polyp with a tubulovillous, but atypical, pattern. The normal samples were macro-dissected, from the same paraffin block as the tumour DNA.

Family	Linkage ID	Type of sample	Tumour ID	Info	Gender	LOH
323	0120_304	TA	0013077_1A	Small	1	loss 1p
323	0120_304	TA	0013077_4A		1	-
323	0120_305	TA	0011426_1A	1.5cm	1	loss 4q, 8q, 15q
323	0120_311	TA	0206460_1A	0.4cm	1	loss 4q
323	0120_301	TVA	04_12868_2C	2.8x1.8x1.5cm	1	loss 15q, 20p
323	0120_301	TVA	04_12868_2A_1	Part 2	1	loss 15q, 20p
323	0120_301	TVA	04_12868_2A_2	Part 3	1	-
323	0120_301	TVA	04_12868_2B	Part 4	1	loss 9q, 15q
323	0120_401	TVA	02_05286_1AT		1	-
336	0122_403	TA	98/685_1A		2	-
336	0122_404	TA	06/009346/3A		1	-
336	0122_301	TVA	38990_B	1cm	1	loss 19q, 4q, 18q
336	0122_301	TVA (atypical)	38990_A	3.5x2.5cm	1	loss 9p, 4q
336	0122_301	TVA (atypical)	38990_E	2.5cm	1	loss 18q, 4q
336	0122_301	TVA (atypical)	38990_C	4cm	1	loss 4q, 18q
336	0122_405	TA	M120505_6_T		1	loss 18q, 19q, 11q
336	0122_405	Normal	M120505_6_N		1	loss 9q CNV?
336	0122_405	Normal	M120505_9_N		1	-
336	0122_405	TA	M120505_9_1		1	loss 11q, 19q
336	0122_405	TA	M120505_9_2		1	loss 19q13.2
336	0122_405	TA	M120505_9_3		1	loss 5q
336	0122_405	TA	M120505_9_4		1	loss 11q
336	0122_405	TA	M163005_B_1		1	-
336	0122_405	TA	M163005_B_2		1	loss 19q, 11q
336	0122_405	TA	M163005_B_3		1	loss 19q, 11q
336	0122_405	TA	M163005_B_4		1	loss 19q
336	0122_405	TA	M163005_B_5		1	loss19q
336	0122_405	TA	M163005_B_6		1	loss 19q, 18q
336	0122_304	TA	02/112441A/C	<8mm	2	-
336	0122_304	TA	02/112441A/D	<8mm	2	-

336	0122_304	TA	02/112441A/F	<8mm	2	-
336	0122_304	TA	02/112441A/H	<8mm	2	-
336	0122_304	TA	02/112441A/I	<8mm	2	-
336	0122_304	TA	03/17024_IAA		2	-
336	0122_304	TA	03/17024 1A		2	-
336	0122_304	TA	07/151131A/A		2	-
336	0122_303	TA	02/11284/2A/D		2	-
336	0122_303	TA	01_07226_1A		2	-
336	0122_303	TA	03_15710_1A	Small	2	-
336	0122_303	TA	0211284_1A		2	-
336	0122_303	TA	0211284_2A		2	-
282	0088_522	TA (rectal)	H004993/C1		2	-
282	0088_522	TA (hepatic)	H004993/A1	2x2mm	2	-
282	0088_522	TA (splenic)	H004993/B1/A	2x3mm	2	-
282	0088_522	TA (splenic)	H004993/B1/B		2	loss 5q
282	0088_407	TA	8546_97_3	2x1x1cm	1	loss 1p
294	0065_301	TVA	97_11282A2_A	1x0.5x0.8cm	2	-
294	0065_301	TVA	97_11282A2_B	Part of above	2	-
294	0065_301	TVA	97_11282B_A	1.1x0.7x0.6cm	2	-
294	0065_301	TVA	97_11282B_B	Part of above	2	-
294	0065_302	TVA	98_41282_A	Split into 3	1	loss 5q
294	0065_302	TVA	98_41282_B	Part 2	1	loss 5q
294	0065_302	TVA	98_41282_C	Part 3	1	loss 5q
294	0065_302	TA	00_8810		1	-
294	0065_302	TA	97_90571_A		1	-
294	0065_304	CRC	9602_1523_3		2	-
294	0065_309	TA	A18_0744100_1T		1	-
294	0065_311	TA	98_1018_A		1	-
450	0162_201	TA	9784/92/1		2	-
450	0162_201	TA	9784/92/2		2	-
450	0162_203	TA	4958/87		2	-
450	0162_103	TVA	91/1551/1		1	-
450	0162_103	TVA	91/1551/2T		1	-
450	0162_103	TVA	99/7772A/A	Part 1	1	-
450	0162_103	TVA	99/7772B/A	Part 2	1	-
329	0109_302	TA	C11008/1/06		2	-
329	0109_302	TVA	C_20642_99		2	loss 18q
377	0125_302	TA	044123_1A_A		2	-
377	0125_302	TA	044123_1A_B		2	loss 1q, 8q, 15q
377	0125_302	TA	044123_1A_C		2	loss 8q, loss 1q, 11q, 18q
377	0125_302	TA	044123_1A_D		2	loss 1q, 17q
377	0125_302	TA	044123_1A_E		2	-
377	0125_302	TA	SS02_037961A		2	-

377	0125_302	TA	SS/02/03796/2A		2	-
377	0125_302	TA	02/037964A/A		2	-
377	0125_302	TA	SS02_037964A_B		2	-
377	0125_302	TA	SS02_037964A_C		2	-
377	0125_302	TA	SS02_037964A_D		2	-
377	0125_201	TVA	848_95		2	-
377	0125_201	TVA	00/03743/3A/A		2	-
377	0125_201	TVA	00/03743/3A/B		2	-
377	0125_201	TVA	00/03743/3A/C		2	-
377	0125_201	TA	00/03743/1A		2	-
377	0125_201	TA	S0110291/1A/A		2	-
377	0125_201	TA	S0110291/1A/C		2	-
377	0125_201	TA	S0110291/1A/D		2	-
377	0125_201	TA	S00_7201_2AA		2	-
377	0125_201	TA	S00_7201_2AB		2	-
377	0125_201	TA	S00_7201_2AC		2	-
377	0125_201	TA	S00_7201_2AD		2	-
377	0125_204	CRC in TVA	97/13626/2A	1.5x1x0.7cm	2	-
377	0125_204	TVA	97/13626/3A	1.5x1x0.7cm	2	-
377	0125_204	TVA	97_13626_1A	2x1cm	2	loss 3q, 5q, 19q
377	0125_205	TA	98_13905_C		2	loss 17q
377	0125_205	TVA	04_06591_E		2	loss 1p, 17q
377	0125_202	TA	99_00083	3mm	2	-
377	0125_203	CRC	95_2882_C_T		1	-
377	0125_203	Normal	95_2882_C_N		1	-

As an example of the data used to call the LOH events listed in the table above, I have included the plots from one of the tumours from the individuals that showed LOH at chromosome 5q (see below, figure 7.4). These were some of the clearest results and many of the others cover smaller regions of LOH. Most of the LOH events were copy neutral and only one sample showed LOH with a clear reduction in copy number. However, this sample was the normal epithelium extracted from the same section as an adenoma in individual 0122_405 (sample M120505_6_N). It is unclear how normal this tissue is as it may have been contaminated by nearby tumour DNA or it may be in the early stages of transformation to an adenoma.

Figure 7.3 B allele frequency and log R ration plots for tumour H005993/B1/B.

The results for chromosome 5 for a tumour (adenoma) from individual 0088_522 of family 282 showing with a clear loss of heterozygotes in the top B allele frequency graph. From the overall pattern of the log R ratio in the bottom plot, there is no indication of a loss in copy number.

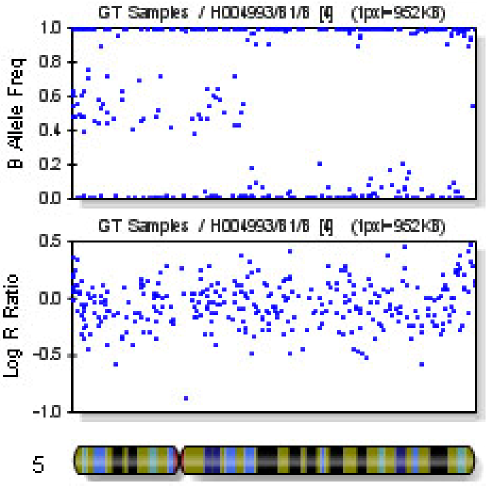
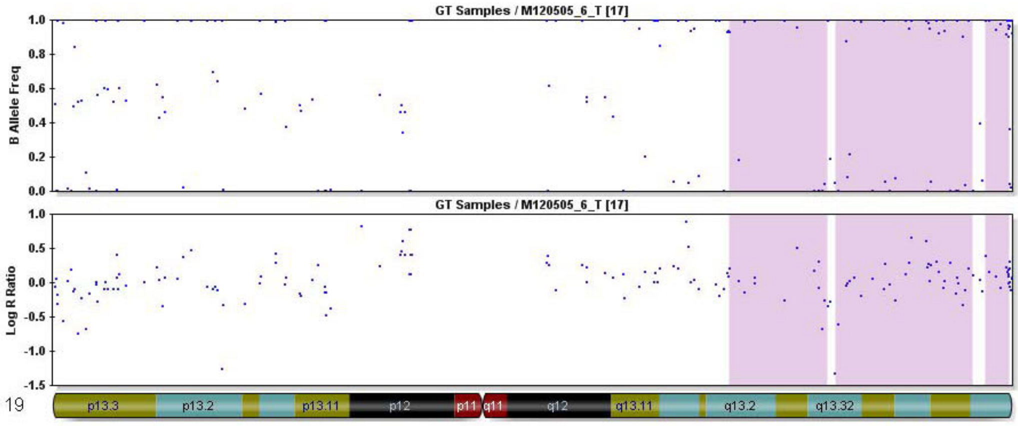


Figure 7.4 The chromosome 19 LOH in family 336 visualised using raw intensity data

Using the Beadstudio B allele frequency and Log R ratio plots, the chromosome 19 region was detected most clearly in this sample (M120505_6_T), which is from 0122_405. Where B allele frequency is 1, all SNPs are homozygous for the B allele. If B allele frequency is zero, the SNPs are homozygous for the A allele. The missing band of points at y=0.5 around 19q13.2 indicate the absence of any heterozygous SNP genotypes, which was confirmed in the genotype data and the ROH analysis. The log R ratio does not show much of a deviation from zero, except at one point (19q13.32) so there does not appear to be a copy number variation affecting this region (copy neutral LOH).



7.3.3 Detection of smaller regions of LOH using homozygosity mapping

As smaller regions of LOH were difficult to determine using just the plots alone, I used a homozygosity mapping approach on the raw data (using the same technique as the ROH analysis described in Chapter 6) and searched the genotypes for continuous runs of homozygous SNPs to determine if there were regions of LOH shared between individuals and within families.

The criteria used for calling a run of consecutive homozygous SNPs as a homozygous segment in the ROH analysis of the tumour DNA, bearing in mind the low density of the linkage SNP panel on which the samples were genotyped, was set to a minimum of 10 consecutive SNPs. I set the minimum length of a segment to 0kb, the density to 5,000kb/snp (i.e. at least 1 SNP every 5,000kb) and maximum allowed gap between SNPs to 5,000kb. I did not make any adjustments to this analysis to allow for the fact that the DNA was extracted from formalin fixed paraffin embedded tissue samples, which are of poorer quality to DNA from blood or fresh frozen tissue. However, individuals with a low genotype calling rate (less than 95%) were excluded from the analysis. Three samples (out of 98) were removed due to poor quality genotyping.

The results from this analysis were compared using the ROH output from PLINK and plotted using the source code for GWA_view (a program written by Jean-Baptiste Cazier, http://www.well.ox.ac.uk/~jcazier/GWA_View.html).

A member from each family was genotyped as part of the EngP1 dataset on the Illumina Hap550 SNP array (see Table 7.5) and the data for ROHs greater than 50 SNPs for these samples (from the homozygosity mapping analysis data, Chapter 6) were used to act as a comparison to the somatic data to determine whether ROHs were germline changes common to the family or individual somatic variations in the tumours. The results for these individuals are plotted along with the tumour DNA in the figures that follow and are labelled 'germline'.

Table 7.5 Individuals from each family that were genotyped as part of the EngP1 dataset

Individuals that were selected from these families for genotyping on the Hap550 SNP arrays were the most severely affected members as defined by phenotype and age of diagnosis. In the results that follow the EngP1 ID has been used to distinguish the tumour DNA sample from the germline DNA.

Family ID	Individual ID	EngP1 ID	Phenotype	Age of diagnosis	SNP-LOH tumour data
336	0122_301	1061D09	9 TA/TVA	43	Yes
323	0120_301	3162A10	1 TVA	35	Yes
377	0125_201	1063F02	43 TA	34	Yes
450	0162_201	1048D04	1 ad	31	Yes
282	0088_409	1050H11	CRC	60	No
294	0065_308	3160G01	1 TA >1cm	46	No
329	0109_303	1063D11	CRC	44	No

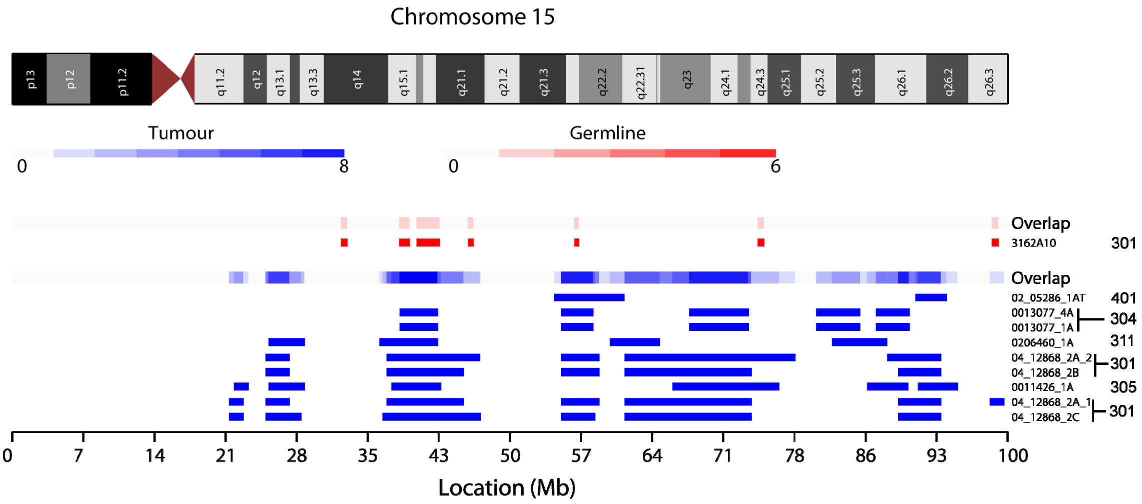
7.3.3.1 Common regions of LOH in the individual families

In the initial screen of the LOH results described above, three families were identified where a number of separate tumours had LOH in the same chromosomal region. Using the ROH analysis of the tumour genotypes, these regions can be identified more easily and additional regions were identified in additional family members. The full details of the individual ROHs comprising the common regions listed below are given in the appendix, Table 9.3.

In the initial results table above, family 323, four tumours showed LOH at 15q (from individuals 0120_305 and 0120_301). Further study of the genotypes for ROHs in all tumours from this family indicated that the common region of LOH was observed in eight tumours from four individuals and has a consensus region of 39,044kb to 42,884kb and includes 12 SNPs covering a 3840kb region (see Figure 7.5).

Figure 7.5 Family 323: chromosome 15 LOH

The ROHs that cover more than 10 SNPs regardless of the overall size are shown. The number alongside the tumour ID is the individual ID. The four adenomas from 0120_301 in this figure (04_12868_2A, 2B and 2C) comprise four sections from the same tubulovillous adenoma.

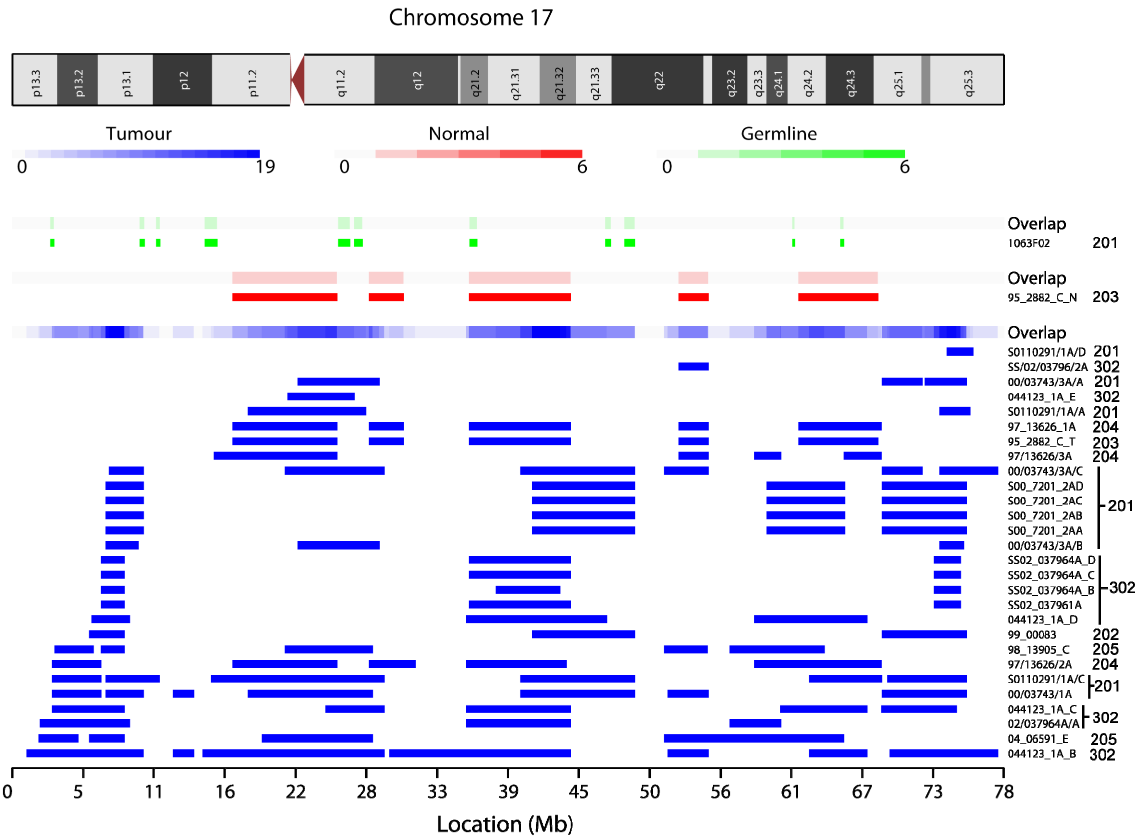


Family 377 showed evidence of LOH at 17q in three tumours from two individuals (0125_205 and 0125_302). A closer look at the genotypes in all tumours from this family, showed a region shared by 20 tumours with a consensus region at 41,302kb and 43,498kb (covering five SNPs and spanning 2197kb, see Figure 7.7, below). LOH at this region is also shared by the normal DNA from 0125_201. The region includes the gene for cell division cycle protein 27, isoform 1 (CDC27), which is part of the anaphase-promoting complex.

The second most common region, which is shared in the 19 tumours, is between 7,708kb and 8,915kb (covering 6 SNPs and spanning 1,207kb) and is not present in the normal DNA sample. Genes in this region include Myosin, heavy chain 10, nonmuscle (*MYH10,* OMIM*:* 160776), which is thought to be involved in ubiquitin-mediated proteolysis required for exit from the cytokinesis phase of the cell cycle, and the gene encoding phosphoinositide-3 kinase, regulatory subunit 5 (PIK3R5), which plays a role in proliferation, cell survival and chemotaxis.

Figure 7.6 Family 377: chromosome 17 LOH

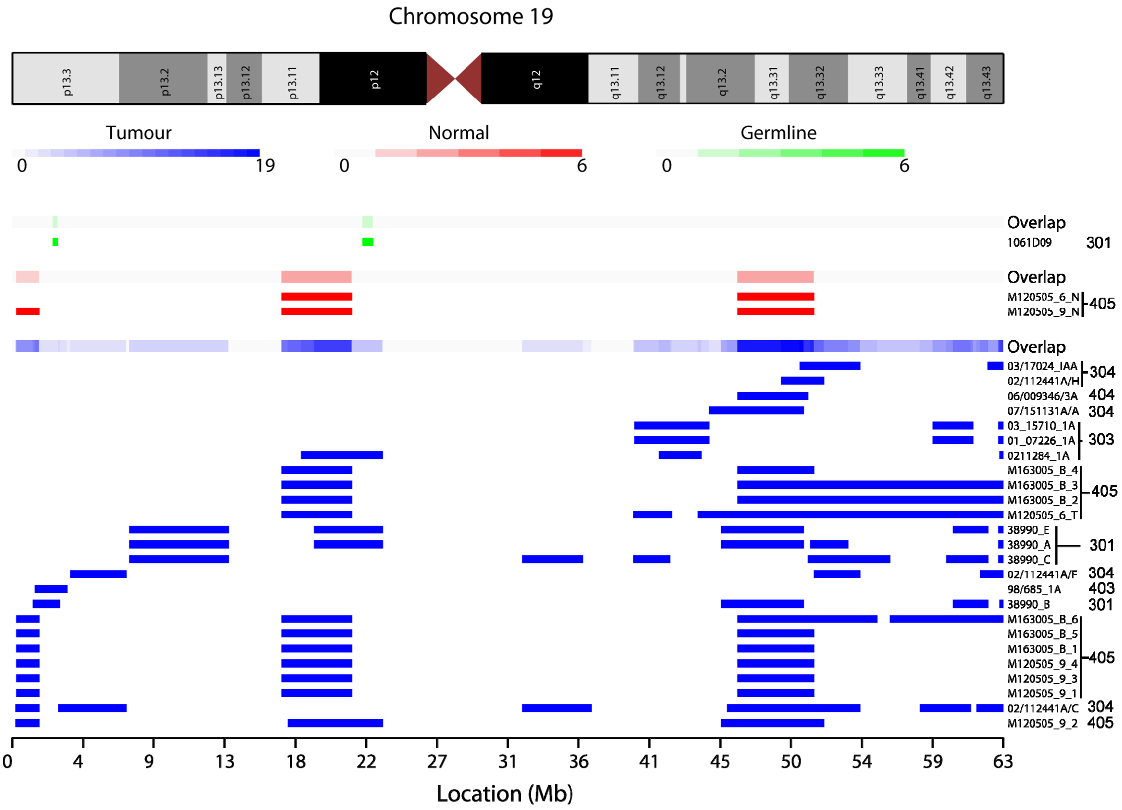
The ROHs that cover more than 10 SNPs regardless of the overall size are shown. The number alongside the tumour ID is the individual ID. The samples from 0125_203 (95_2882_C_T and C_N) were extracted from the same tumour block. The similarity in the patterns of LOH between these two sections suggests some contamination of the normal tissue with tumour DNA, although the pattern is also shared by 0125_204.



Family 336 had multiple tumours with LOH at 19q (see Figure 7.8, below). The most common region of LOH (pool S154, see appendix) was between 50,683kb to 50,930kb and is present in 21 tumours from four individuals (0122_301, 0122_405, 0122_304 and 0122_404). However, most of the tumours belong to two individuals 0122_405 and 0122_301, who are father and son. The ROHs detected in the germline DNA of individual 0122_301 do not overlap with this region, suggesting the presence of heterozygous genotypes indicating that a somatic alteration has occurred. No regions of linkage were identified in this region. There are a number of genes in this region, including vasodilator stimulated phosphoprotein (*VASP*), which encodes a protein associated with filamentous actin formation that plays a role in cell adhesion and motility, and the gene encoding gastric inhibitory polypeptide receptor (GIPR), which is expressed by K cells located in the duodenum and small intestine, which inhibits the production of gastric acid and promotes insulin secretion.

Figure 7.7 The LOH region on chromosome 19 for family 336 as defined by the presence of ROHs

This figure includes all ROHs detected in this family covering more than 10 SNPs regardless of overall size. The tumour 98/685_1A is from 0122_403 (5 from the bottom), who does not carry a ROH near the region common in the other samples. The germline DNA is from 0122_301, genotyped on the Illumina Hap550. The DNA for the normal samples is from individual 0122_405 and was extracted by macro-dissection from the same paraffin block as the tumour DNA. This figure illustrates that four cases (0122_304, 405, 301 and 304) out of the six studied have ROHs overlapping the same region.



7.3.3.2 LOH in regions common in the tumour progression pathway

The results of this analysis indentified a number of interesting regions of LOH that do not map to regions of detected linkage. There are several regions of LOH that coincide with the locations of known tumour suppressor genes, such as APC on chromosome 5 (at 112Mb), SMAD4 on chromosome 18 (at 46.8Mb) and TP53 on chromosome 17 (7.5Mb). These events are commonly seen in colorectal cancers and are important steps in the adenoma to carcinoma progression pathway (Vogelstein *et al.* 1988; Fearon and Vogelstein 1990), described in Section 1.2.1. As the majority of the tumours included in this study are adenomas, I did not expect to see a large number of LOH events at these locations and there were no samples with large regions of LOH (whole chromosome arm or a substantial proportion of a chromosome arm) across all three of these regions.

There were, however, large regions of LOH on the q arm of chromosome 5, detected in three adenomas from 0065_302 (family 294), which were all sections from the same TVA sample, and one adenoma from 0088_522 (family 282). The locations of these regions were clear from the ROH results and show LOH spanning the location of the *APC* gene. In Family 294, this region is not shared by the other tumours screened, although other smaller regions were detected (see Figure 7.9 and the Table 9.3 in the Appendix for the precise locations of these regions). The consensus LOH region (S47) shared by 11 adenomas in this family spans 1,439kb between 79-80Mb and contains the gene for muts homolog 3 (MSH3), which is part of the DNA mismatch repair system. Mutations in genes involved in miss-match repair can cause MSI and have been described in relation to HNPCC and endometrial cancer.

The single region of LOH identified in the initial scan of the adenomas from 0088_522, is actually shared by four adenomas belonging to the same individual and extends from 97.99 to 118.6Mb, which includes *APC* (see Figure 7.9).

Figure 7.8 The chromosome 5 region in Family 294

The ROH results of chromosome 5, which highlights the LOH region identified from the intensity data shown above. The three tumours showing the largest regions of LOH belong to the same individual (0065_302), but few of the other family members share a large region of LOH in same location. The samples 98_41282_A, B and C are three sections from the same TVA. Equally, samples 97_11282B_A and B and samples 97_11282A2_A and B from 0065_301 are both bisected adenomas. The germline (blood DNA) results are from individual 0065_308, who did not have adenomas included in the analysis. However, only relatively short ROHs can be seen on this higher density SNP array (Hap550 panel); the region is broken up by heterozygous SNPs.

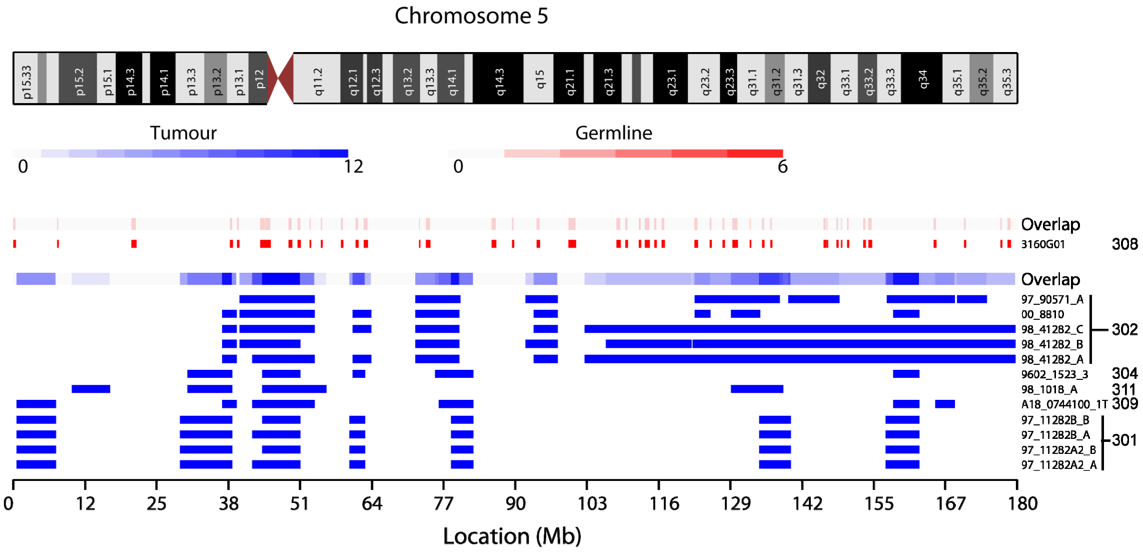
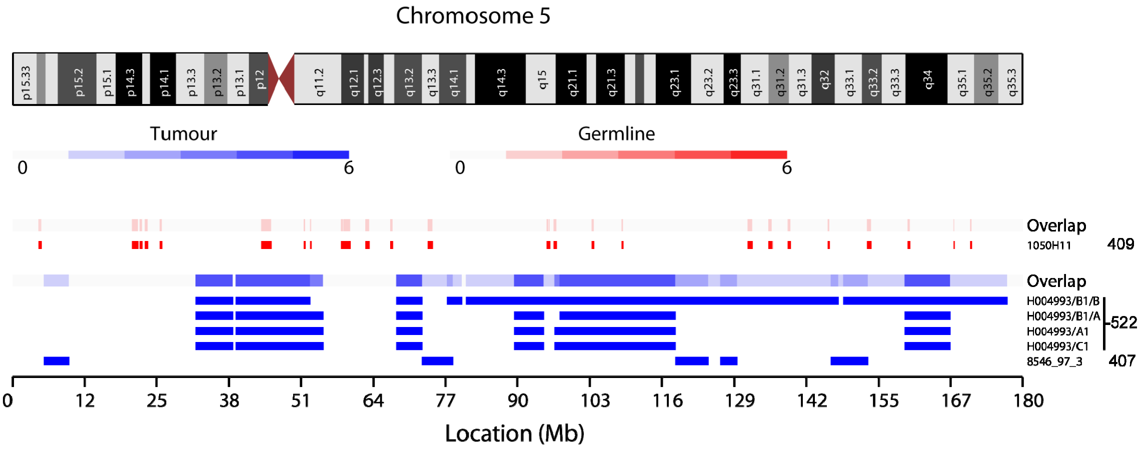


Figure 7.9 The chromosome 5 LOH region in family 282

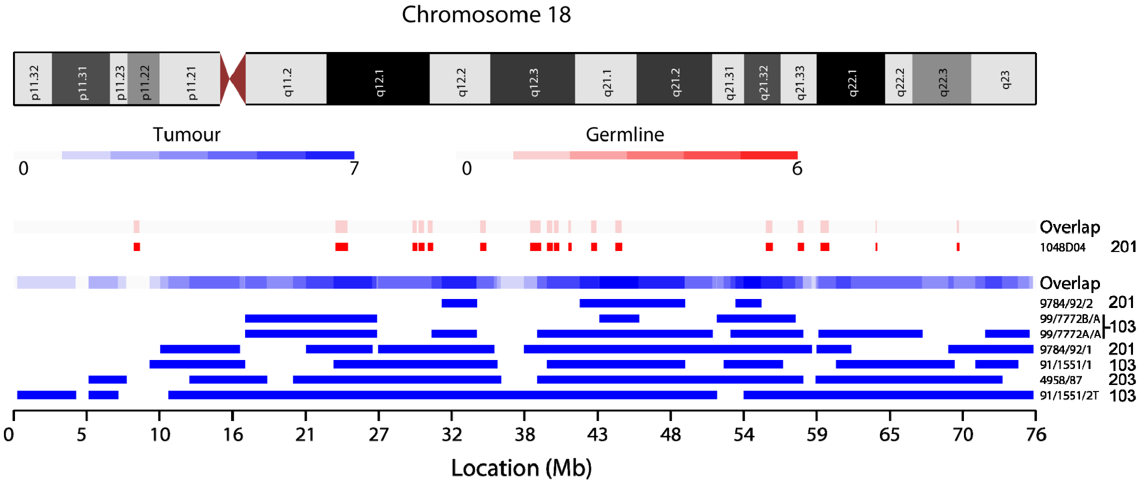
The large region of LOH on chromosome 5 in 0088_522 and the three additional separate adenomas showing smaller regions of LOH also belong to the same individual. The germline DNA is from individual 0088_409, whose tumours were not available for LOH analysis.



Additionally, in family 450, I identified LOH on chromosome 18 in all seven tumours from three individuals analysed (3 TAs and 4 TVAs, see Figure 7.10). The samples 99/7772A/A and 99/7772B/A are two parts of the same TVA. Six of these samples showed LOH spanning the location of SMAD4 at 46.8Mb and, incidentally, SMAD7 at 44.7Mb, which was identified as part of the GWA study (consensus region of 43.6Mb to 49.9Mb). However, in tumour 99/7772B/A the LOH region only reaches to 46.5Mb. Juvenile polyposis can be caused by mutations in *SMAD4*, although Juvenile polyps do not appear to be common in this family. However, on the basis of these results, this family was screened for *SMAD4* mutations as it was discovered that this possibility has not been ruled out previously through clinical genetic testing. A 4bp duplication resulting in a frame-shift in *SMAD4* has now been discovered in two individuals in this family.

Figure 7.10 The LOH region on chr18 in family 450

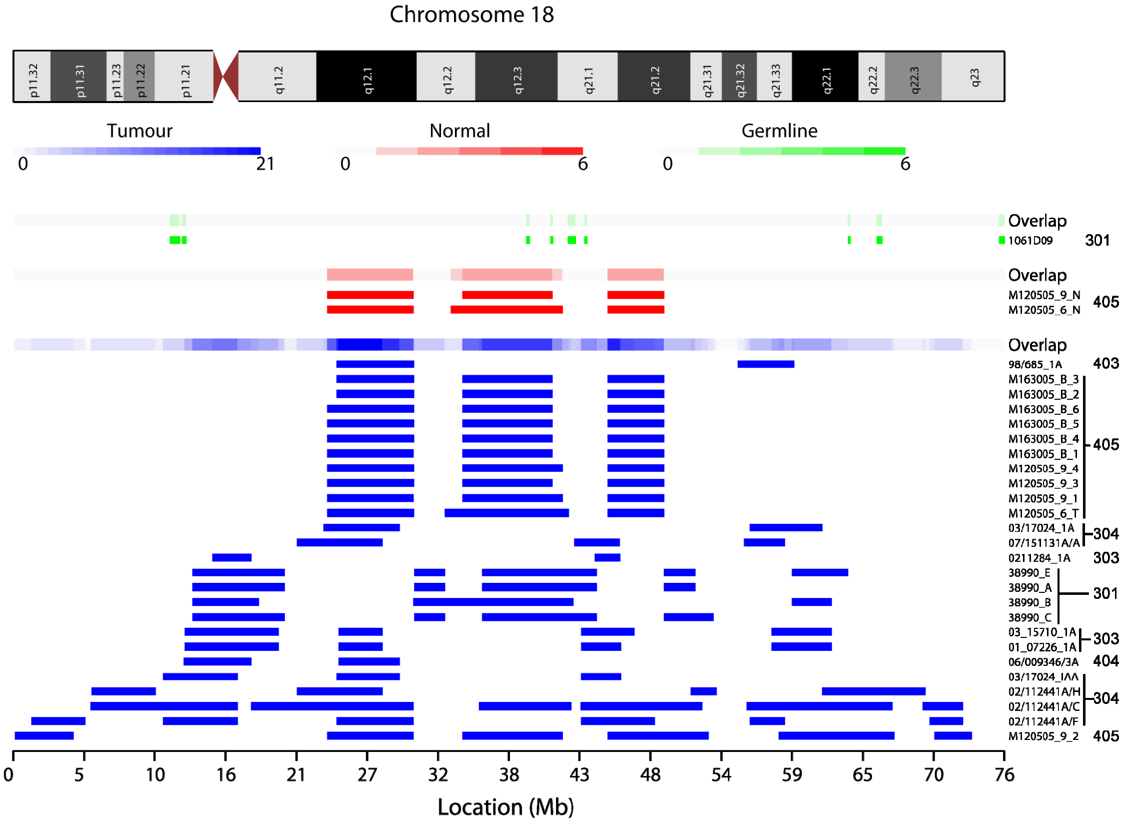
The presence of LOH at this region of chromosome 18 is interesting because it contains *SMAD4* and is a common site of LOH in colorectal tumours and one of the changes that occurs in the progression of adenoma to cancer. All 3 members of this family have tumours showing LOH in this region. The germline (blood) DNA is from 0162_201. The samples 99/7772B/A and 99/7772A/A from 0162_103 are two sections of the same TVA.



LOH at 18q was also identified in five samples from family 336 (Figure 7.11, below), but although this region included SMAD7 (which is an antagonist of transforming growth factor-beta (TGF-β) signalling and linked to CRC), the LOH region did not cover the SMAD4 gene in five tumour samples (the consensus region spanned 45.6Mb to 46.5Mb, see appendix Table 9.3) and this region did not show LOH in 0122_301.

Figure 7.11 The chromosome 18 LOH region in Family 336

Small regions of LOH were detected in 6 individuals in this family, but the consensus region considering all samples (45.6-46.5Mb) did not cover *SMAD4* (due to five of the tumours), although it did include *SMAD7*, which is linked to CRC.



Two other samples from 329 and 377 also showed LOH on chr18, although this was not more common in the TVAs compared to the TAs. I did not detect any large regions of LOH at 17p, but as this is generally seen later on in tumorigenesis, and there were only two cancers in the dataset, this was not unexpected. There were a number of regions that were common among all tumours studied, but these were generally quite small (less than 2Mb, see Table 7.6).

Table 7.6 The four most common regions across all tumours analysed

The consensus regions of the ROH's detected in the analysis that comprised tumour samples from the highest number of individuals across all families analysed is given. However, these consensus regions are actually very small, only covering two or three SNPs.

Pool	No of tumours	Chr	Start (SNP)	End (SNP)	Start (bp)	End (bp)	Size (kb)	SNP
S4	71	6	rs508557	rs6915493	75374741	75688699	313.958	2
S3	71	22	rs760519	rs1534880	35588206	35653611	65.405	2
S9	69	15	rs872263	rs2047415	72923497	74167595	1244.1	3
S11	69	3	rs11714798	rs1392695	97892561	98368882	476.321	2

A summary of the results from the LOH analysis, for each family, which gives an overview of regions of all ROHs greater than 4Mb in size detected in each family, is given as a genome-wide summary in the Appendix figures 9.7-9.12.

7.3.3.3 Comparison of regions of LOH with the detected linkage peaks

Using the overlapping regions of the detected ROHs, I compared the ROH locations in each family with those of the linkage peaks described above. ROHs overlapped with regions of linkage for three regions detected in family 336, three in family 377 and one in family 294. These regions of LOH (as defined by ROHs) were detected in a number of tumours within each family and could indicate the presence of tumour suppressor genes (see Table 7.7). However, LOH in the region on chromosome 10 in family 336 is only seen in the adenomas from 0122_301.

The LOH regions identified are relatively large and include many genes that could be involved in susceptibility to cancer, including transcription factors and regulators of the cell cycle. For example, in family 336, the consensus LOH region on chromosome 3 at 161,287kb to 162,849kb includes the gene for structural maintenance of chromosomes protein 4 (SMC4), which is involved in DNA repair. The chromosome 3 region in family 377 covers 127.6-133.1Mb and includes the gene that encodes the GATA binding protein 2 (GATA2), which is a transcription factor important in the regulation of development and proliferation in haematopoietic and endocrine cells and WD repeat-containing protein 10 (WDR10), which has roles in cell cycle progression and apoptosis. The region on Chromosome 10 detected in family 336 contains the gene for Catenin alpha-3 (*CTNNA3*), which has been reported to recruit beta-catenin and E-cadherin (CDH1) and to medate cell-cell adhesion (Janssens *et al.* 2001).

Table 7.7 Summary of Linkage peaks with LOD>1.4 and comparison with detected LOH regions

This table shows the boundaries of the linkage peaks (defined by LOD>1) detected in the recessive and dominant linkage analyses, the value and location of the maximum LOD scores and any runs of homozygous SNPs detected in individuals of the same family that are located in the region of the linkage peaks, as defined by the consensus region of the ROH positions. Consensus regions (con) were calculated using the region where the most tumours carry overlapping ROHs. The data for the samples and ROHs contributing to each of these pooled ROHs is given in the appendix along with the union of all the ROHs contributing to each pool (Figure 9.4 appendix). The ROH pool S1663 on chr. 10 for family 336 only contain the four tumours from 0122_301 and so is it difficult to draw conclusions about its significance in relation to the linkage peak. In all of the detected linkage peaks all affected members of the family shared the risk haplotype, except the region between 22 and 25Mb on chromosome 3 in family 336, where the haplotype was not shared by 0122_405.

Family	Linkage region		Peak (kb)	Max LOD	Chr.	Summary of consensus LOH regions							
	From (bp)	To (bp)				Pool	No of tumours	SNP (start)	SNP (end)	Start (kb)	End (kb)	Size (kb)	SNPs
	Dominant Model												
336	45,262,492	47,462,139	47,228	1.66	7	-	-	-	-	-	-	-	-
	61.24cM	64.44cM	64.15cM										
336	60,710,738	72,381,709	67,304	1.72	10	S1663	4	rs911610	rs1227938	64,290	70,828	6538.23	10
	75.65cM	90.43cM	81.82cM										
	Recessive Model												
336	33,769,570	37,802,755	34,071-34,397	1.67	2	-	-	-	-	-	-	-	-
	55.32cM	61.07cM	56.0-56.5cM										
336	220,480,390	231,111,135	229,348	1.69	2	S146	22	rs1431087	rs997363	226,309	228,335	2025.62	4
	217.81cM	237.40cM	232.16cM			S11	29	rs375154	rs936070	225,705	226,209	503.801	3
377	193,149,195	217,541,644	207,657-210,969	1.41	2	S228	17	rs896441	rs7014	195,092	196,351	1259.17	3
	191.50cM	213.27cM	203.2-205.3cM			S2019	2	rs2715896	rs869134	201,267	201,384	117.688	2
336	22,549,173	25,899,702	24,028	1.66	3	-	-	-	-	-	-	-	-
	41.88cM	46.79cM	43.93cM										
336	150,868,229	170,440,142	161,951-165,654	1.67	3	S496	16	rs12634498	rs4305435	161,287	162,849	1562.11	5
	156.1cM	168.8cM	164.7-165.7cM			S1119	8	rs755763	rs1388007	153,483	154,739	1256.05	3

						S1232	7	rs9438	rs6773566	155,501	156,507	1005.43	4
377	116,802,290	130,208,720	117,753-121,846	1.41	3	S350	15	rs1799404	rs6792114	127,641	133,018	5376.76	14
	121.25cM	135.05cM	122.29-125.19cM			S1524	5	rs1127343	rs634265	123,611	125,587	1975.7	5
						S2011	2	rs1147696	rs1472621	121,602	123,495	1893.25	5
282	653,347	13,806,353	7,358-7,957	1.46	3	-	-	-	-	-	-	-	-
	0cM	31.39cM	19.12-20.69cM										
294	11,773,330	28,142,830	23,712-24,064	2.65	7	S16	12	rs6463843	rs1029718	8,612	15,817	7205.58	15
	1.13cM	1.02cM	34.42-34.94cM										
377	13,041,712	27,218,503	24,912	1.41	9	S1626	4	rs1128957	rs560764	25,667	29,583	3915.58	10
	26.71cM	49.55cM	46.4-48.1cM			S217	17	rs7866589	rs1328273	14,238	16,013	1775.68	5
366	14,379,011	16,806,655	15,745	1.67	9	-	-	-	-	-	-	-	-
	30.24cM	35.22cM	32.85cM										
329	104,130,302	106,774,958	105,941	1.63	12	-	-	-	-	-	-	-	-
	7.22cM	122.38cM	121.2-122.0cM										

7.3.4 Discussion of linkage and LOH results

The linkage results from the families analysed separately highlight the heterogeneity between families and the difficulty in detecting causal mutations in families with a severe phenotype and a pattern of inheritance closely resembling an incomplete penetrance, dominant model. It is possible that most of the families analysed do not have quite the number of affected individuals needed to achieve the required power to detect a disease segregating allele. Additionally, it is difficult to confidently assign unaffected status in situations of late onset disease and, therefore, currently unaffected family members were classed as unknown.

Family 336 showed the highest number of regions segregating with disease, with four regions showing a LOD score greater than 1.4. There are a number of interesting points about this family, such as the existence of three members with endometrial cancer and a young age of diagnosis with significant adenomas, especially in the youngest generation. This has the hallmarks of HNPCC, but no mutations have been discovered. The family pedigree suggests a dominant mode of inheritance and indeed the maximum LOD score achieved was 1.72 under a dominant model for the region on chromosome 10, which is very close to the maximum estimated LOD score (1.79). However, although a region of LOH was identified at the same locus, the four tumours involved belong to just one patient (0122_301).

Many of peaks detected in the linkage analysis were in similar locations to regions of detected LOH in the tumours of affected individuals, perhaps indicating the presence of a TS gene. However, although there are many potential candidate genes in each region, it is difficult from this analysis to draw anything more conclusive from these results. Clearly, there are several promising candidate regions that would benefit from being sequenced and further work is required to elucidate the causal genes in these families. However, whole genome, or exome, sequencing of each family member is probably required to truly pin down the genetic susceptibility to CRC.

With regard to the determination of LOH through runs of homozygous SNPs in tumour samples, the small number of overlapping SNPs between the 10K linkage panel used for the genotyping of the tumours and the Hap550 SNP array (on which the germline DNA was genotyped), made it difficult to say whether the ROHs detected in the tumours are true regions of LOH. Owing to the low density of the 10K linkage panel, there may be unobserved heterozygous SNPs located in between the homozygous SNPs in this panel. From the data presented here, it would appear that many of the large ROHs detected in the tumours are not due to an inherited germline region of homozygosity, but have arisen from genetic alterations in the tumour itself. As tumour DNA is very limited in these samples and many samples only contained enough DNA for this experiment, it would be difficult to genotype this DNA on a more dense panel of SNPs. However, it would be very informative to genotype the germline DNA for all individuals used in this study on the same SNP panel as the tumours, to allow a comparison of the differences between the normal and tumour DNA. This was attempted with the normal tissue that was macro-dissected from the same slides as three of the tumours, but this was not available for all samples and there is an increased chance of contamination from the adjacent tumour tissue making it difficult to be sure how genetically ‘normal’ this tissue really is.

There were a number of common regions of LOH within families that were not common amongst all the families studied, although these regions were not identified in the linkage analysis. The results from family 450 led to the family being clinically screened for mutations in *SMAD4* and frameshift mutations have been identified in two individuals so far. Although the elusiveness of the causal factors in these families is frustrating, the results of this analysis has identified several additional candidate regions for further research into the susceptibility variants causing disease in each of these families, which will ultimately aid their future clinical management, but work on these families is continuing.

7.4 Analysis of regions shared identical by descent

7.4.1 Introduction

As a complementary method to the association analysis, I have performed a population-based linkage analysis using GWA data to detect shared regions among cases and test whether these segments are associated with disease risk (Purcell *et al.* 2007). Essentially, distantly related individuals that are affected with the same disease could provide additional information for gene mapping of the disease susceptibility locus.

The rationale behind this experiment is that disease susceptibility may be caused by multiple rare variants. If there are several independent rare variants in the same gene or region (not tagged by the same common variant) that individually explain increased risk in only a few cases in the population studied, they will not be detected by GWA methods. However, GWA data can be used to analyse the sharing of genomic segments between cases in an approach similar to linkage. One would expect susceptibility variants to reside in regions that are shared between pairs of cases more often than between pairs of controls. This technique analyses shared segments rather than frequencies of a single SNP or haplotype and therefore should provide a complementary approach to detecting susceptibility variants in a population based linkage-like study. This analysis consists of three steps (described in the Materials and Methods, Section 2.13): determining the level of relatedness between individuals by identifying shared segments identical by state (IBS), use this data to estimate segments identical by descent (IBD) and then analysing these pairs of samples for phenotype correlations.

7.4.2 Detecting segmental sharing between individuals

The SNPs used in this analysis are in approximate linkage equilibrium and the panel was chosen by pruning the SNPs in the Hap550 array based on a pair-wise r^2 threshold of 0.2. If there is LD between SNPs, the shared segments will be longer and it may lead to inflated test statistics and potentially spurious associations.

The first step is to determine the relatedness between individuals by identifying segments shared IBS using the genotype data for independent SNPs. The segmental sharing IBD is estimated from the IBS sharing using a Hidden Markov Model (HMM). The output of this analysis provides a list of all shared regions between individuals with sample ID, physical position and size of region. The total size of shared regions between any two individuals is also calculated and common regions of shared segments are given where each shared region (found in more than one sample) is grouped into pools providing the number of cases and controls included in each pool.

This method has been used to detect shared segments IBD in the HapMap Phase II samples to estimate the degree of relatedness among seemingly unrelated individuals (The International HapMap Consortium *et al.* 2007).

7.4.3 Statistical analysis of shared segments IBD

The final step is to perform an analysis of the shared segments IBD to identify segments that are shared more often between cases than controls. This was performed in PLINK, by labelling pairs into groups of case/case, case/control and control/control and analysing using a one sided test whether there is a higher rate of sharing in case/case pairs compared to control/control or discordant pairs.

It is important to note that the samples included in each pair of shared segments are not independent as each sample may be a member of multiple pairings. The analysis is performed over 10,000 permutations by randomly switching the phenotype labels of the individuals to produce empirical significance values. These are defined by the number of shared segments that span each SNP to which the significance value is assigned. Therefore, each result is not independent, as a segment may span several SNPs.

Any significant results will identify a region of the genome that contains significantly more segments shared IBD in case/case sample pairs than in other sample pairs. However, this test statistic is still in development and so any results must be considered with a degree of caution.

7.4.4 The Jewish samples from the EngP1 dataset

The samples included in this analysis were identified as an outlying cluster in the PCA analysis described in section 3.3.1.2 and found to be of mainly Jewish descent. Susceptibility variants for CRC in the Ashkenazi Jewish population have been discovered previously and the strong relatedness in these samples, compared to our largely outbred north European datasets, may facilitate the detection of additional susceptibility variants. This group of samples consists of 31 cases of which 18 are affected with CRC and 13 are only affected with significant adenomas (22 males and 9 females) and 24 controls (12 males and 12 females). As there was sufficient numbers of cases and controls in this dataset, the opportunity arose to perform a small analysis in this population. Therefore, I decided to perform a linkage type analysis to look for shared regions that were estimated identical by descent (IBD) and identify any differences between cases and controls that could affect disease risk in this population. The phenotype of the cases used in this analysis is described in Table 7.8.

Table 7.8 Phenotype details of the 31 cases included in this analysis

This table describes the cases that were included in this analysis and details whether the individual was affected with cancer or adenomas or both. The sex is labelled 1 for males and 2 for females. The Dukes stage is given where this information was available.

ID	Sex	Cancer (Y/N)	Dukes stage	Adenomas (Y/N)	Total No adenomas	Age of diagnosis
1048F06	1	Yes	-	N	-	<70
1050F12	1	Yes	A	Y	-	68
1053B06	2	Yes	-	N	-	25
1053C05	2	Yes	-	N	-	45
1053D04	1	Yes	-	N	-	20
1053H06	1	Yes	-	N	-	49
1055H09	2	No	-	Y	5	57
1059A07	2	Yes	B	N	-	35
1059B04	2	No	-	Y	6	50
1061C12	1	Yes	-	N	-	50
1061D11	1	No	-	Y	1	38
1061G05	1	Yes	-	N	-	49
1061H12	2	Yes	A	N	-	70
1062A07	2	No	-	Y	3	70
1062B06	2	Yes	B	Y	7	25
1062C02	1	Yes	C	N	-	48
1062C06	1	Yes	A	Y	8	37
1062D02	1	No	-	Y	3	56
1062G02	2	No	-	Y	3	60
1062G07	1	No	-	Y	4	68
1062H02	1	Yes	C	N	-	48

1062H12	1	No	-	Y	1	35
1063A07	1	Yes	C	N	-	67
1063G08	1	Yes	B	N	-	75
3160A07	1	No	-	Y	4	39
3160D07	1	Yes	A	N	-	53
3162A10	1	No	-	Y	1	35
3162B10	1	No	-	Y	13	53
3162D02	1	No	-	Y	5	59
3162F11	1	No	-	Y	7	65
3162G12	1	Yes	C	Y	3	65

7.4.4.1 The results of the analysis

There were 80,400 SNPs in the pruned SNP panel that were in approximate linkage equilibrium (based on a pair-wise r^2 threshold of 0.2). The results of this analysis were based on 3,317 shared segments that meet the size criteria, of which 1,485 pairs of shared segments were shared between affected individuals. As the segments used to generate statistics in this analysis are not independent (a sample sharing a segment in one pair may also share the segment with other samples and, as in linkage analysis, the results assign a P value to each SNP that may be incorporated in many different segments), a Bonferroni correction, as used in single SNP association studies, to determine a formally significant P value would be too stringent. Therefore, I have taken a significance threshold equivalent to that used in linkage analyses (LOD>3.3) and used a P value less than $1x10^{-3}$ as formally significant. The genome wide results of this analysis indicate that chromosome 5 and chromosome 12 contains regions that are suggestive of an association with CRC risk (Figure 7.12) and I have discussed the results for these chromosomes separately below.

Figure 7.12 A genome wide summary of the results of the IBD association analysis in the EngP1 Jewish samples

This plot shows a genome-wide overview of the results of the association analysis of shared segments between pairs of samples. The P value is a one sided test of segment sharing in case/case pairs.

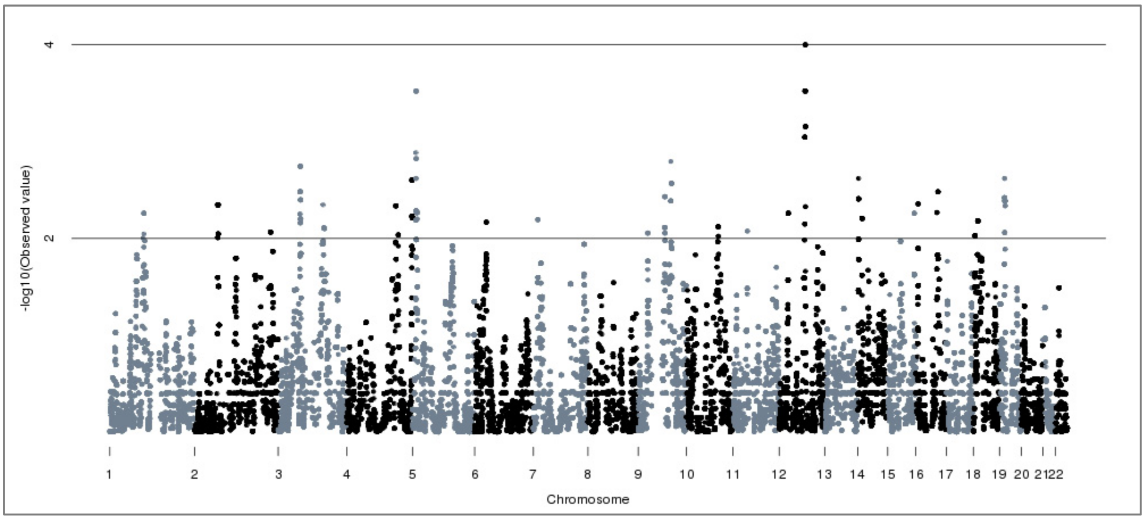
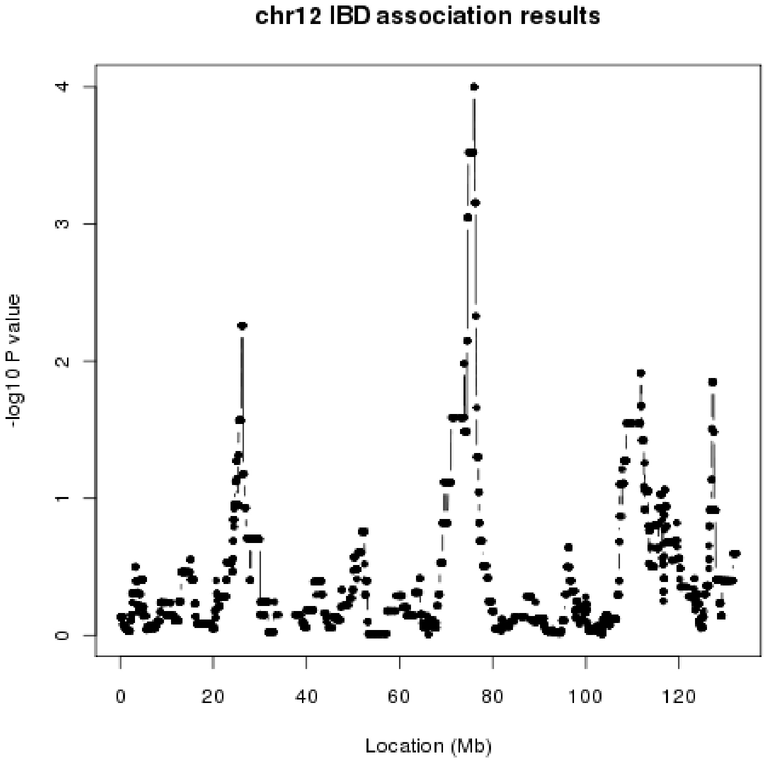


Figure 7.13 Shared segments IBD Association Results for chromosome 12

The empirical significance values for the case/case shared segments in the Jewish dataset. The peak (spanned by 8 case/case pairs of segments) is located at approximately 76Mb, with a P value of 9.99x10^{-5}.

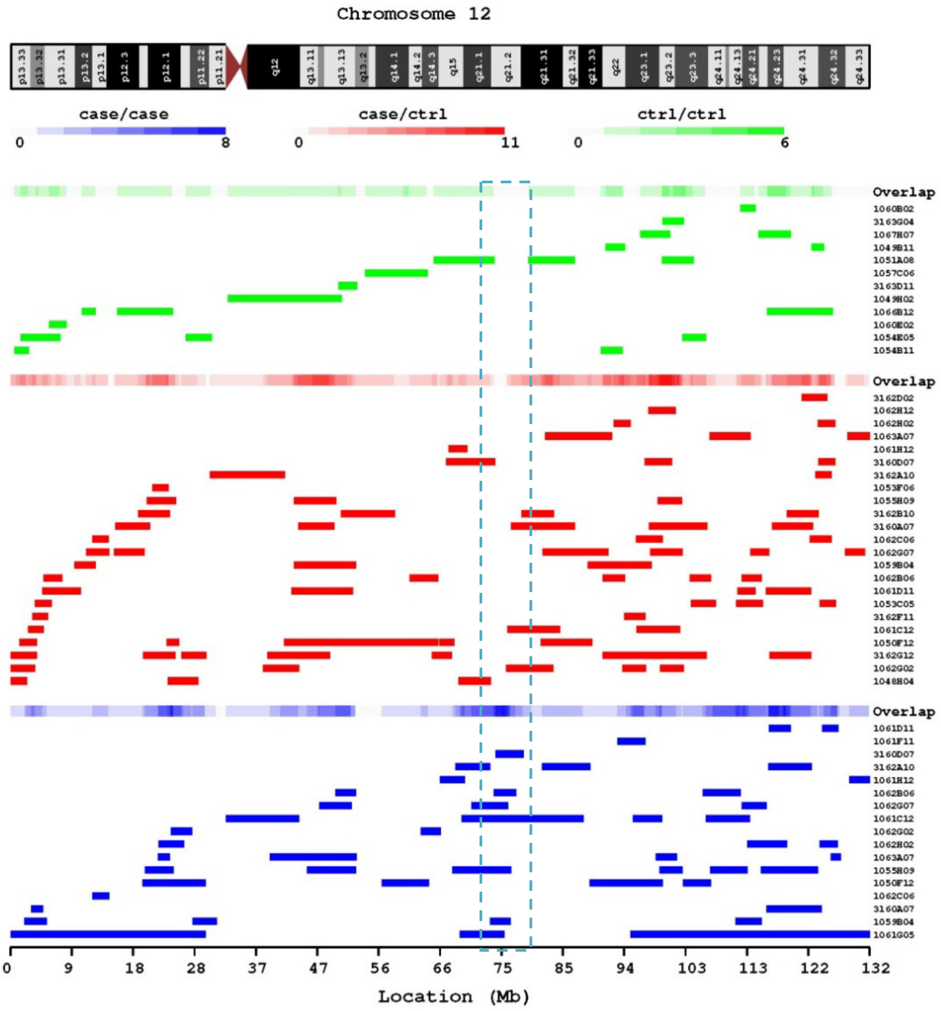


The most significant peak on chromosome 12 spans from 73,872,453 (rs2446346) to 76,390,931bp (rs2203995), with a minimum P value of 9.99x10^{-5} for the region spanning 75,921,829 (rs310886) to 76,033,220bp (rs10506731, see Figure 7.13). The locus is spanned by 8 segments shared IBD in case/case pairs. No non-case/case pairs shared a segment in this locus (see figure 7.14). This region contains several genes including E2F7, located at 75,939,157-75,983,49bp, which is an E2F transcription factor that is an essential regulator of the cell cycle by the regulation of genes whose products are needed for cell cycle progression (Di Stefano *et al.* 2003) and NAP1L1 (nucleosome assembly protein 1-like 1), located at 74,726,223-74,764,717bp, this gene encodes a protein that plays a role in DNA replication and the regulation of cell proliferation. The expression of the protein is increased in rapidly proliferating cells in mice (Hajkova *et al.* 2008).

Also in this region are the genes for OSBPL8 (oxysterol-binding protein-like protein 8 isoform), which is located at 75,269,709-75,477,720bp, is a member of an intracellular lipid receptor family and BBS10 (Bardet-Biedl syndrome 10) located at 75,262,397-75,266,353bp, which has roles in cilia formation and mutations in this gene cause Bardet-Biedl syndrome. This condition is characterised by progressive retinal degeneration, obesity, polydactyly, renal malformation and mental retardation (Stoetzel *et al.* 2006).

Figure 7.14 The location of shared IBD segments on chromosome 12

The locations and size of the all pairs of shared segments between case/case, case/control and control/control pairs of individuals. The highest point in the peak was located at about 76Mb, the approximate region is marked with a hashed box, and was based on segments shared IBD in 8 case/case pairs and no case/control or control/control pairs.

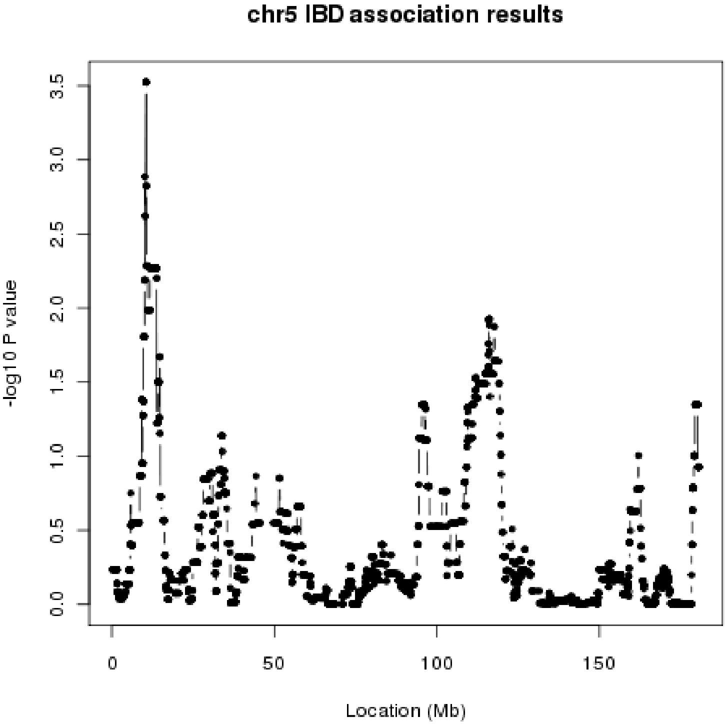


The next highest peak was located on chromosome 5 spanning 9,506,089bp (rs436243) to 14,857,735bp (rs4702054, see Figure 7.15) with a peak at 10,525,134bp (rs7711645) -10,621,417bp (rs2399910) with a P value of 3.00x10^{-3}. This result is based on seven shared segments between case/case pairs and no non-case/case pairs. The distribution of all shared segments detected on this chromosome is given in Figure 7.16.

The region contains, among others, the gene Catenin (cadherin associated protein) delta 2 (CTNND2), which promotes the disruption of CDH1 (E-cadherin) and is over expressed in prostate adenocarcinoma.

Figure 7.15 IBD Association Results for chromosome five

The empirical significance values for the case/case shared segments in the Jewish dataset. The peak spans 9,506,089bp to 14,857,735bp with a P value of 3.00x10^{-3} and is shared between 7 case/case sample pairs and 0 control/control samples at about 10.5Mb.

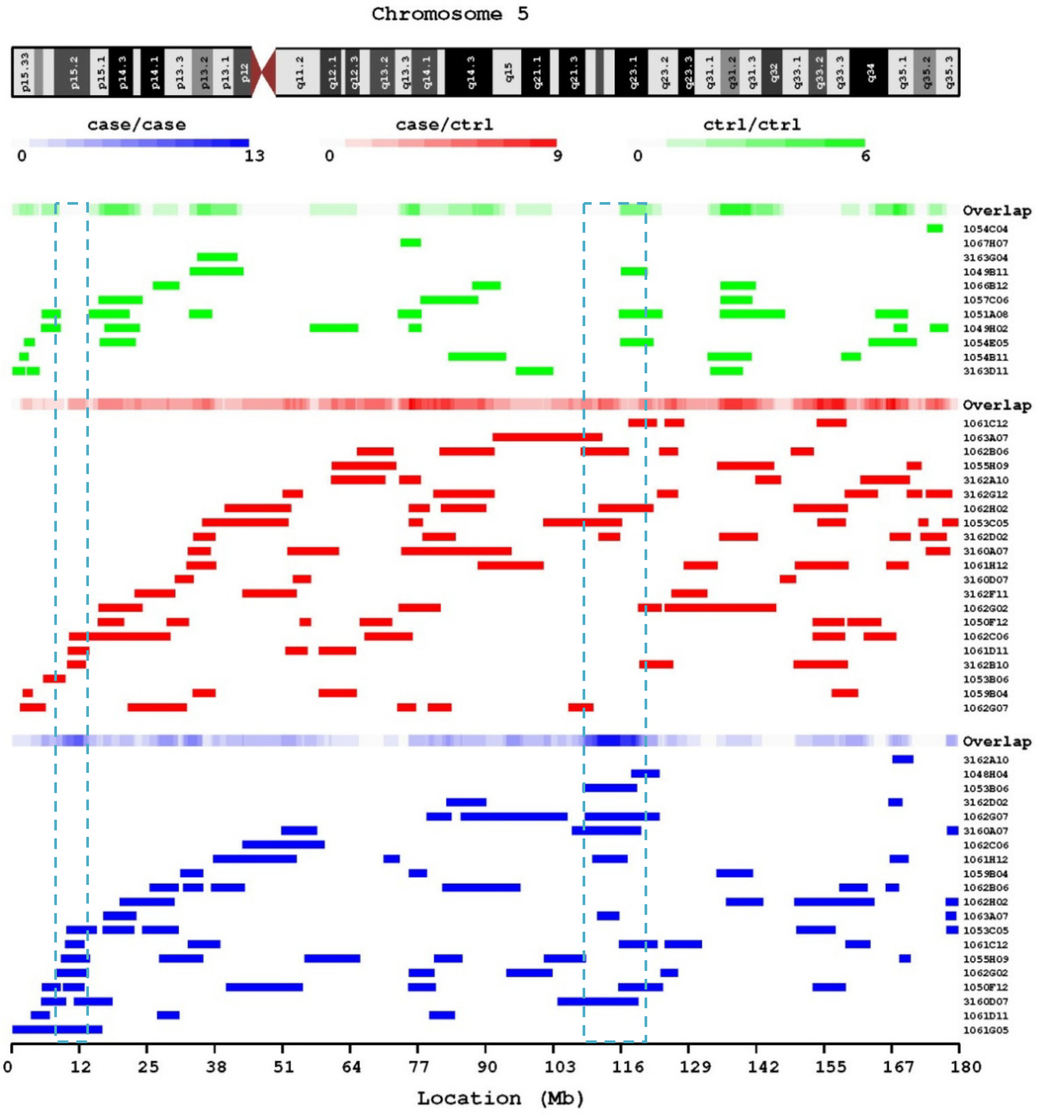


There is an additional interesting region on this chromosome that spans 109,258,634 (rs245243) to 119,695,731bp (rs10051178) with a peak at 116,002,152bp (P=0.0119, rs2112655) to 116,188,771bp (rs153577) and incorporates 12 concordant case/case, 3 discordant and 1 control/control pair of shared segments IBD. The region includes a number of genes including *APC*, which is a tumour suppressor gene that negatively regulates the Wnt signalling pathway (mutations in this gene cause FAP), and mutated in colorectal cancers (*MCC*), which has two isoforms and is thought to negatively regulate the cell cycle. This region also contains the calcium/calmodulin dependent protein kinase IV (*CAMPK*), which is involved in transcriptional activation in lymphocytes and neurons and the gene small conductance calcium activated potassium (*KCNN2*), which encodes an integral membrane protein that forms part of a calcium activated potassium channel.

However, this result could be an indication that the cases harbour the *APC* variant, I1307K, which was discovered in individuals from the Ashkenazi Jewish population (Laken *et al.* 1997) where it appears with a frequency of approximately 5% and has been shown to confer at 1.5 fold increase in CRC risk (Cazier and Tomlinson 2009). Work is ongoing in the dataset.

Figure 7.16 The location of detected segments IBD on chromosome 5

The locations and size of the all pairs of shared segments between case/case, case/control and control/control pairs of individuals. The most significant point is shared in 7 case/case pairs of individuals at about 10.5Mb, the second highest peak is a region shared in 12 case/case pairs and located at about 110Mb. The approximate regions are marked with hashed boxes.



7.4.5 Analysis of segments shared IBD in the Scottish dataset, ScotP1

The ScotP1 dataset is made up of Scottish cases and controls and owing to the smaller area in which these samples were recruited there is the possibility that this dataset is less outbred than the other datasets included in the study. In order to assess the possibility of enrichment on the basis of shared regions of the genome between cases, I performed an IBD analysis, as above. The results of this analysis were then compared in cases and controls in an association like test, as above, to determine whether there are regions that are shared IBD more often between pairs of cases than pairs of controls or case/control pairs.

This analysis was based on 965 cases and 984 controls over 103,871 approximately independent SNPs, based on pair-wise r^2 using a threshold of 0.2. Pair-wise r^2 between SNPs was calculated using the ScotP1 samples included in this study (using the pruning function in PLINK, as described previously). In this analysis there were a total of 557,549 segments detected that met the size criteria, of which 142,097 segments were shared between pairs of affected samples. As previously, I have used a P value approximately equivalent to a LOD score of 3.3 in a linkage analysis and taken $P<1x10^{-3}$ as being significant. A genome wide summary of the results is given in Figure 7.17.

Figure 7.17 IBD association analysis results for ScotP1

This plot shows a genome-wide overview of the results of the association analysis of shared segments between pairs of samples. The P value is a one sided test of segment sharing in case/case pairs.

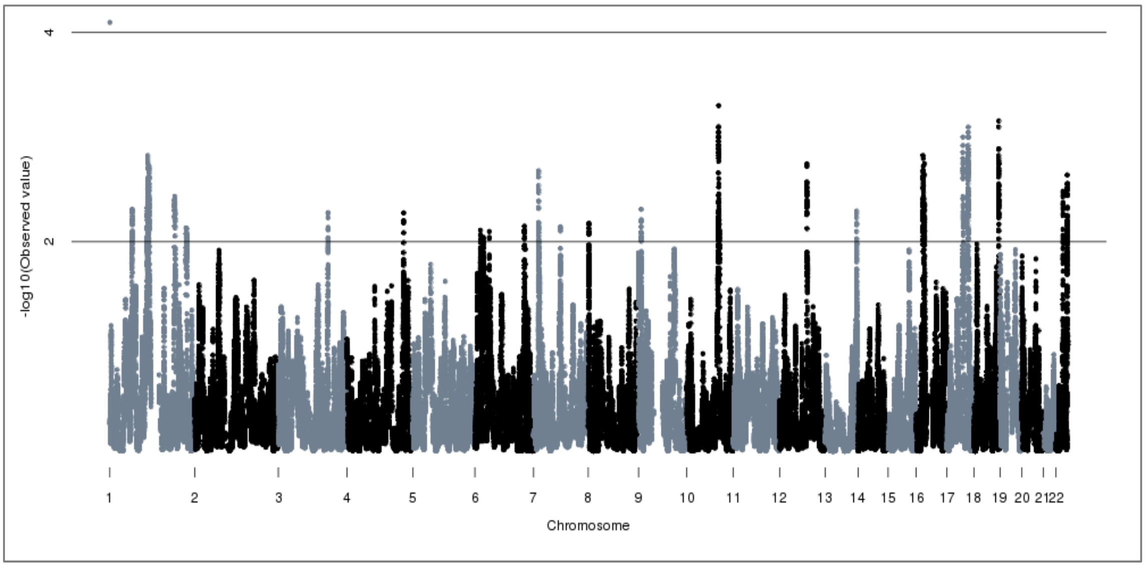


Table 7.9 The summary of the IBD analysis results

The most strongly associated segments that are shared IBD. The shared segment is represented by a SNP that is located within each of the included segments. It is important to note that the samples in the case/case, case/control and control/control groups may not be independent and this is accounted for by performing permutations of the sharing calculations while swapping the individual's phenotypes. The P value (EMP1) is a one sided test of increased case/case segment sharing and relates to the analysis of the segments shared IBD that cover the listed SNP and not the actual SNP. The positions in this table are genome build 35. The start and end positions show the boundaries of the peak defined by segments with empirical significance P values less than 0.1 either side of the lowest P value.

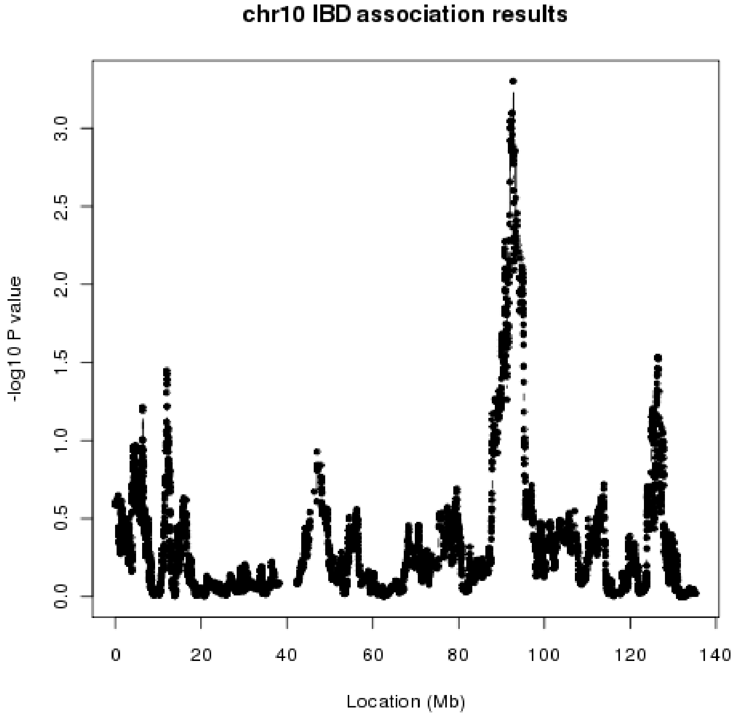
Chr.	Lowest P value	SNP covered by shared IBD segment	Position (bp)	Start (bp, SNP)	End (bp, SNP)	Number of segment pairs shared IBD			Interesting genes in the region under the association peak
						case/ case	case/ control	control/ control	
10	5.00x10^{-4}	rs747334	92,734,724	89,154,174 rs12251529	95,293,024 rs10882265	224	333	141	PTEN, FAS, ANKRD1, PP1R3C
17	9.99x10^{-4}	rs16950116	47,022,532	45,734,614 rs7225245	50,715,263 rs3829577	236	339	163	NME1 TOB1
17	7.99x10^{-4}	rs7224730	62,746,699	60,820,652 rs9891078	64,920,257 rs8071378	156	208	112	AXIN2 PRKCA
18	6.99x10^{-4}	rs7238069	72,591,405	71,869,875 rs1395813	73,181,529 rs948709	159	219	107	MBP, ZNF516, ZNF236, GALR1

The chromosomes with the strongest associations (10, 17 and 18) are plotted separately in Figures 7.18, 7.19 and 7.20 and the most strongly significant results are given in Table 7.9. Owing to the large number of samples in this study it was not practical to include a graphical representation of the shared IBD segments as provided in the section above, but the numbers of shared segments that fall into each group are given in the table. The individuals in the shared segments are not necessarily independent and an individual in one pair may also be included in another pair.

The most significant loci on chromosome 10 (see Figure 7.18) spans a number of interesting genes in relation to the potential to influence CRC risk. These include the tumour suppressor gene phosphatase and tensin homolog (*PTEN*), which is mutated in a high number of cancers and negatively regulates the AKT/PKB signalling pathway, the gene that encodes a receptor belonging to the tumour necrosis factor receptor superfamily member 6 (*FAS*), which contains a death domain and is central to the control of programmed cell death and its mutation is associated with various cancers, and the cardiac ankyrin repeat protein (*ANKRD1*), which functions as a tumour suppressor and is induced by TNF-α and IL-1.

Figure 7.18 IBD association results for chr10

This plot shows the empirical significance value, generated from the analysis of greater segment sharing between cases, for each loci (SNP) spanned by a shared IBD segment on chromosome 10.

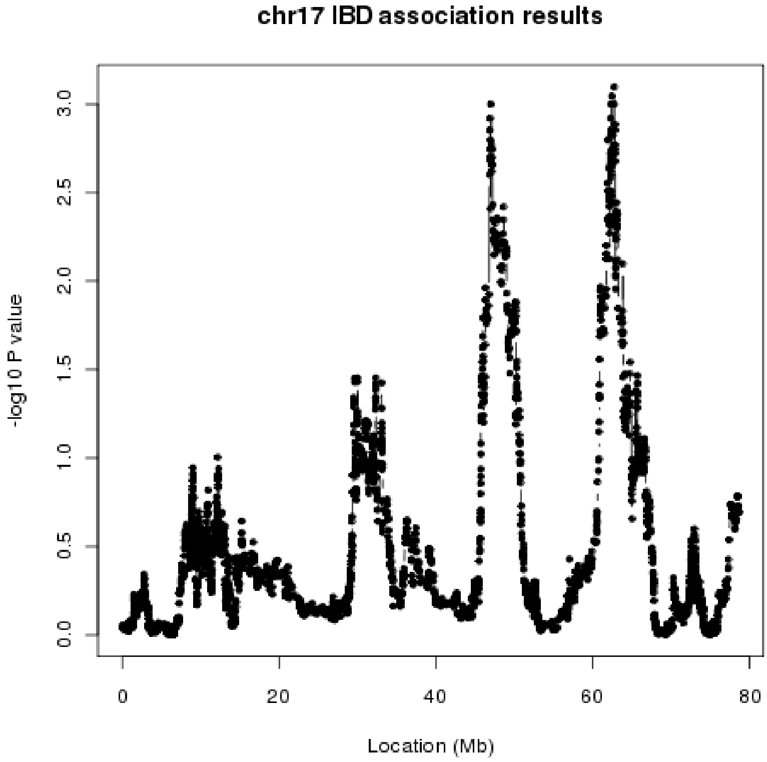


The results for chromosome 17 (see Figure 7.19, below) actually show two peaks of similar significance, the first at about 47Mb. This locus contains the genes nucleoside diphosphate kinase 1 (*NME1*), which shows decreased expression in highly metastatic cells and is commonly mutated in aggressive neuroblastomas, and transducer of ERBB2 1 (*TOB1*), which is part of a family of proteins that negatively regulate cell growth. The protein encoded by this gene interacts with SMAD2 and SMAD4 to enhance their activity and has been shown to inhibit T cell proliferation and transcription in cytokines and cyclins.

The second peak on chromosome 17 is at approximately 63Mb (60.8-64.9Mb), which spans the genes axin regulated protein (*AXIN2*), which controls the stability of β-catenin as part of the Wnt-signalling pathway. The locus is a common site of LOH in cancers such as breast cancer and neuroblastoma, and protein kinase C alpha (*PRKCA*), which encodes a protein with an important role in cell signalling involved in cell adhesion, transformation and cell cycle checkpoint activation by calcium and diaglycerol.

Figure 7.19 IBD association results for chr17

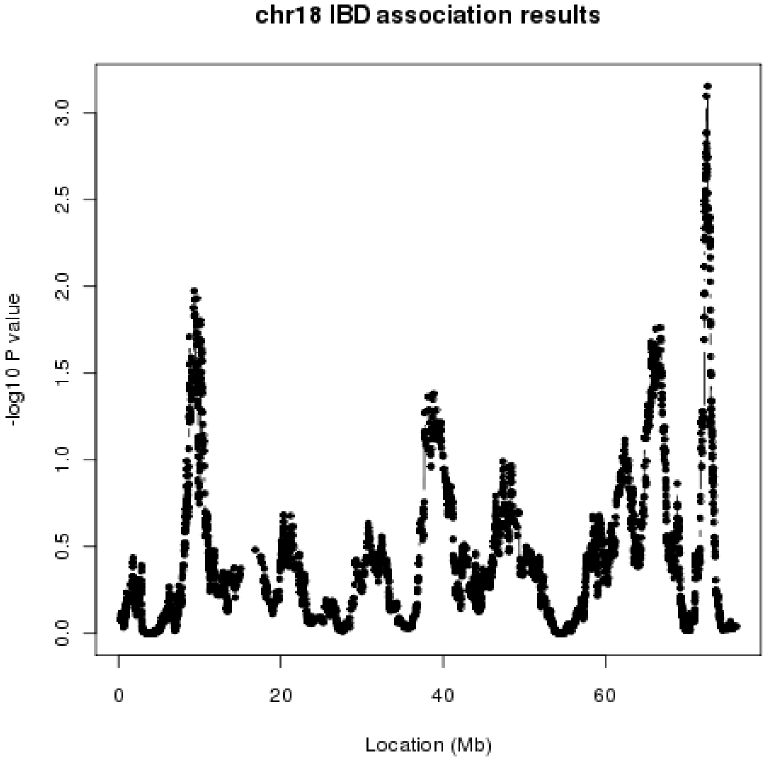
This plot shows the empirical significance value, generated from the analysis of greater segment sharing between cases, for each loci (SNP) spanned by a shared IBD segment on chromosome 17.



The peak on chromosome 18 (Figure 7.20) is located at approximately 72.5Mb (71.8-73.2Mb) and contains the genes Myelin basic protein (*MBP*, Golli-mbp isoform1), the zinc finger proteins: ZNF516 and ZNF236, and the gene encoding the neuropeptide galanin receptor 1 (*GALR1*), which is a G protein coupled receptor that inhibits the action of adenylyl cyclase and is expressed in the small intestine, among other sites.

Figure 7.20 IBD association results for chr18

This plot shows the empirical significance value, generated from the analysis of greater segment sharing between cases, for each loci (SNP) spanned by a shared IBD segment on chromosome 18.



7.5 Discussion of the IBD association analysis results

Although this analysis has not been widely used in the literature and the analysis described here is a preliminary one, the results shown indicate that the method is promising as a genome-wide multipoint method for the detection of regions associated with disease.

Although there was not much evidence of this technique being validated on other GWA study datasets, the fact that most of the genes present within the loci defined by the detected peaks are functionally relevant to influence CRC susceptibility supports the value of this method and this was especially evident in the ScotP1 analysis. The regions detected do not largely coincide with regions identified previously using the single SNP GWA approach and require further investigation. Developments in this technique and in particular the accuracy of inferring IBD have been made recently by another group (Bercovici *et al.* 2010) and it may be informative to repeat the analysis to compare the results.

The first analysis in the Jewish samples was under powered and the P values of most associated segments rely on the segment sharing in only a handful of samples. However, this group of samples is likely to be less distantly related than the ScotP1 samples and hence there may be a higher chance of detecting shared segments that are enriched for the disease allele.

Both analyses identified loci that reach a genome-wide significance level, based on a linkage threshold of $1x10^{-3}$ and it will be interesting to see if these regions can be replicated in independent larger studies. However, it was difficult in this study to determine the threshold for significance of the detected regions and how best to correct for multiple testing when the segments were not independent. As this analysis was more akin to a linkage analysis, I used a standard linkage significance threshold, which may not be appropriate for this test. However, this has not resulted in implausible regions of association and with the relatively small numbers of samples (especially in the Jewish dataset analysis) I did not expect to achieve statistically significant results. I have also not made any adjustments for the uncertainty in statistically estimating IBD from the IBS data in the absence of parental information.

The method certainly holds promise as a complementary method to GWA studies for the detection of susceptibility variants, although further work is required, and ongoing to determine whether this is successful and refine the analysis.

Chapter 8. Conclusion

The results of the GWA study described in this thesis have shown, by the detection and verification of 14 independent associated SNPs, that common variants do indeed influence the risk of CRC, mitigating the CDCV hypothesis. However, the effect size of these associations has been small and, even in combination the discovered risk alleles do not fully explain the missing heritability of this disease. This validates the view that missing heritability for complex diseases will not be discovered by just one approach and that it is probably a combination of the CDCV and CDRV hypotheses, plus additional alterations such as copy number variants, that will explain the total variation in risk. In a similar way to how linkage studies showed that the total variance in risk could not be explained by highly penetrant, rare variants, the GWA era has shown that it cannot be fully explained by low penetrance, common variants either.

GWA studies have identified hundreds of novel associated variants that have been robustly replicated in independent datasets in numerous diseases. This study has shown that additional variants of smaller effects can be detected through meta-analysis with additional cohorts to increase sample sizes, especially if all the cohorts have been genotyped on similar whole-genome arrays rather than just the most strongly associated SNPs from the discovery phase.

The results have also enabled further understanding of the pathways and cellular functions involved in disease risk. Four of the 14 associated SNPs identified for CRC tag genes that function in the BMP signalling pathway, which suggests that disruption of this pathway is important in CRC predisposition. However, genes such as *CDH1* and *ATF-1*, act in multiple pathways, which could all play a part in disease susceptibility.

It is clear that there is still plenty of work to be done in the GWA datasets described in this thesis. For example, further analysis in increased numbers of samples of the X chromosome variants to validate the two identified SNPs and also the additional methods explored in this thesis, such as homozygosity mapping and pair-wise segmental sharing analyses of distantly related individuals, could yet yield additional susceptibility variants. As shown in Chapter 7, large families of affected individuals exist where no Mendelian mutation has been discovered to explain the increase in CRC risk. Through familial approaches and genome-wide or exome sequencing, these individuals could be used to aid the detection of rare variants that are likely to be enriched amongst affected members, but could also aid the detection of causal mutations in families with a near Mendelian disease inheritance. LOH studies and the search for somatic mutations in the early tumours of patients is valuable when compared with germline DNA to better understand the pathways involved in tumorigenesis and can aid the identification of causal mutations. The results in one family with LOH on 18q led to the screening of the *SMAD4* gene and the identification of a deleterious mutation in two individuals in that family.

8.1 Where do we go from here in the detection of missing heritability for common complex diseases?

The addition of more samples will only take us so far in the detection of small effect common variants and many questions remain unanswered. Are the identified significant common variants actually tagging rare variants that are the true causal alleles? How much missing heritability is being overlooked due to the problem of heterogeneity at the disease locus, as is common in Mendelian diseases such as FAP, causing low relative risks or leading to variants not being detected? Are real associations going unnoticed due to the, perhaps necessarily, stringent genome-wide significance threshold? GWA studies have suffered problems with prioritisation of associated SNPs, as most SNPs identified for replication studies have generally come from those most associated in the original discovery phase. Although this captures the most significant SNPs, this method is likely to have missed those SNPs that are truly associated with disease, but are rare or have a small effect size and so do not reach the threshold for significance.

Despite the encouraging GWA study results, critics of the CDCV approach cite the limitations of using indirect GWA methods, based on HapMap tagging SNPs, such as poor representation of SNPs with low minor allele frequency and the possibility of allelic heterogeneity at the disease locus, as reasons for its perceived inadequacy. One downside of the GWA study approach is that although we have discovered many common variants associated with disease susceptibility, the method of using tagging SNPs has meant that few causal variants have been directly typed and, hence, they are difficult to identify. Therefore, with the increasing accessibility of technology for sequencing large numbers of samples, research is already moving in the direction of rare variants. The identification of rare variants (and the genotyping of all common variants), through next generation sequencing efforts may help elucidate the true causal variants behind the association signals as well as uncovering additional factors that may influence disease susceptibility.

The study of rare variants could identify susceptibility alleles that have not been discovered in a GWA approach owing to heterogeneity at the disease locus, where multiple independent rare variants in different individuals affect the same gene. This problem has been approached in analyses of rare variants in rheumatoid arthritis using whole-genome association data and a program called GRANVIL (Morris *et al.* 2009). The authors developed a method to identify genes that may harbour a number of rare susceptibility variants that in combination confer a small increased risk, which would not be identified in a single SNP analysis.

Additionally, using the already genotyped GWA cohorts, a free approach would be to impute the additional common variants (MAF greater than 1%) in the latest release of the 1000 genomes project. The subsequent large meta-analysis for association could greatly improve the fine-mapping of detected regions and better define sites for future sequencing efforts to determine causal alleles.

Determining causal variants should enable identification of the genes affected and, hence, a better understanding of the cellular functions and pathways that influence CRC risk. This could help identify regulatory factors involved and identify affected genes, as seen with the 8q24 SNP rs6983267 which was not predicted to influence the gene *c-myc*, but was later discovered to affect the expression of a regulatory element that interacts with this gene. This may be the case with other SNPs and until we understand better the regulatory mechanisms controlling the effects of these genes the functional effects of the causal variants will be difficult to identify.

Equally, the LD between tagging SNPs and causal variants will rarely be perfect and so the estimated relative risks attributed to the detected common variants may be much lower than that of the actual causal variant. This could be the reason for lower than expected estimates of the heritability attributed to the identified genetic variants.

The CDRV hypothesis needs to be explored to determine its validity; however, there are difficulties to be addressed with the detection and analysis of rare variants. The most obvious is the need to sequence a very large number of individuals in order to determine statistically significant association with disease and this is still an expensive undertaking. Additionally, although the risk estimates of the identified common variants have been precise with narrow 95% CI, estimates of the relative risk associated with rare variants have been less exact with very wide confidence intervals. For example, the *BRIP1* risk variant in breast cancer with a 95% CI of 1.59-73.4 (Rahman *et al.* 2007).

Prioritisation of SNPs for follow up studies may also present an issue with whole-genome or exome sequencing as the variants taken forward will likely be those chosen based on existing evidence, such as GWA studies.

Furthermore, there are a number of additional avenues to explore in relation to CRC risk determination to include gene-gene interactions, gene-environment interactions and chromosomal aberrations, such as copy number variation. Of course, more functional investigations are required into the effects of candidate genes in associated SNP regions to determine the relevance of genetic changes identified by these methods.

8.2 Overall implications of the research results

The aim of this research was to identify low penetrance common variants that influence the risk of CRC. The identification of susceptibility variants that explain the increased familial risk could form a panel of SNPs for individual risk prediction. This could eventually be combined with clinical data and used to improve the clinical management and prophylactic treatment of patients through earlier detection of individuals at risk of developing the disease, but there is some way to go before this goal becomes a reality.

In terms of clinical relevance for predicting an individual's risk of disease and influencing prophylactic treatment, the actual causal variant is not required if it is sufficiently tagged by the identified variants, assuming that enough of the variance in risk can be explained by these variants. At present, the number of detected susceptibility variants carried by an individual is not conclusive, or complete, enough to determine that individual's risk of developing CRC. The associated risk alleles are common in the population and confer a relatively small increased risk of disease and the functional effect of many of the associated loci is unknown. Added to this, genetic testing alone to ascertain risk of complex diseases is limited owing to environmental interactions that influence disease susceptibility and are unlikely to reach an accuracy that would be clinically useful in the general population. The genetic results, when combined with clinical and lifestyle data, will enable better stratification of high risk individuals for improved personalised clinical management. Despite this, companies such as deCODEme and 23andMe (using 8 and 3 SNPs, respectively) have prematurely jumped at the chance to sell tests of genetic risk for CRC, and other complex diseases, to the public.

The NHS CRC screening program is now fully rolled out in the UK with the aim to send faecal occult blood tests every two years to individuals between the age of 60 and 75 years (http://www.cancerscreening.nhs.uk/bowel/index.html). Abnormal tests results are then followed up with further tests or colonoscopy to help improve the early detection of tumours and, thus, the chance of successful treatment. The pilot study successfully detected 552 cancers of which 48% were Dukes stage A and 1% had metastasised. Eventually, it may be possible to use susceptibility variants, coupled with clinical data, to determine individuals that might benefit from undergoing this type of screening at an earlier age. However, this is not yet a reality and individuals without a known family history of the disease are unlikely to request genetic testing.

From a research perspective, the results of this study has identified loci and genes associated with complex disease that increase our understanding of the molecular pathways and processes involved in tumorigenesis and uncover potential drug targets. Many of these loci were not previously linked to CRC or had not been generally linked to cancer and provide data for pathway analyses and studies into downstream interactions not previously considered important to cancer development.

In summary, the continual wave in the literature of validated variants associated with complex diseases that are discovered by GWA studies speaks for itself of the success of the method in achieving what it set out to do, to identify common variants associated with disease risk using the genome-wide indirect approach of tagging SNPs. When GWA efforts for complex diseases began it was not financially or technically feasible to search for rare variants to explore the CDRV hypothesis and it is very easy to criticise the GWA approach for its shortcomings when the alternative is largely untested. It is important at this stage that researchers from both camps are open to the pros and cons of both approaches as they should be utilised in a complimentary manner for the detection of risk alleles and the realisation of a common goal.

Large scale sequencing efforts for the detection of additional rare and common variants and the fine mapping of associated loci are the current direction for many studies. Hopefully, this will also uncover causal variants to explain the identified associations, provide increased understanding into complex disease susceptibility and allow focussed functional studies on candidate genes.

Supporting Publications

For the papers listed here I published in my maiden name of Spain.

A ‘*’ indicates joint contribution to the work

Houlston, R. S., J. Cheadle, Dobbins, S. E., Tenesa, A, Jones, A. M., Howarth, K, Spain, S. L., *et al*. (2010)."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 **42**(11): 973-7.

Spain, S. L., J. B. Cazier, R. Houlston, L. Carvajal-Carmona and I. Tomlinson (2009) Colorectal cancer risk is not associated with increased levels of homozygosity in a population from the United Kingdom. Cancer Res. **69**, 7422-9.

Houlston, R. S., E. Webb, P. Broderick, A.M. Pittman, M.C. DiBernardo, S. Lubbe, I. Chandler, J. Vijayakrishnan*,* K. Sullivan, S. Penengar, L. Carvajal-Carmona, K. Howarth, E. Jaeger, S. L. Spain *et al.* (2008)."Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer.” NatGenet **40**(12):1426-35.

Pittman, A.M., E. Webb, L.Carvajal-Carmona, K. Howarth, M.C. DiBernardo, P. Broderick, S. Spain, A.Walther*, et al.* (2008). "Refinement of the basis and impact of common 11q23.1 variation to the risk of developing colorectal cancer.” HumMolGenet **17**(23):3720-7.

Jaeger, E.*, E. Webb*, K. Howarth*, L. Carvajal-Carmona*, A. Rowan*, P. Broderick*, A. Walther*, S. Spain**, et al.* (2008). "Common genetic variants at the CRAC1 (HMPS) locus on chromosome 15q13.3 influence colorectal cancer risk." NatGenet **40**(1):26-8.

Tomlinson, I. P., E. Webb, L. Carvajal-Carmona, P. Broderick, K. Howarth, A. M. Pittman, S. Spain, S. Lubbe*, et al.* (2008). "A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3." NatGenet **40**(5):623-30.

Broderick, P., L. Carvajal-Carmona, A.M. Pittman, E. Webb, K. Howarth, A. Rowan, S. Lubbe, S. Spain*, et al.* (2007). "A genome-wide association study shows that common alleles of SMAD7 influence colorectal cancer risk." NatGenet **39**(11):1315-7.

Tomlinson, I., E. Webb*, L. Carvajal-Carmona*, P. Broderick*, Z. Kemp*, S. Spain*, S. Penegar, I. Chandler*, et al.* (2007)."A genome-wide association scan of tagSNPs identifies a susceptibility variant for colorectal cancer at 8q24.21." NatGenet **39**(8):984-8.

References

Websites for programs and tools used in this thesis

1000 Genomes Project http://browser.1000genomes.org/index.html

GLIDERS http://mather.well.ox.ac.uk/GLIDERS/

GWA_view http://www.well.ox.ac.uk/~jcazier/GWA_View.html

GTOOL http://www.well.ox.ac.uk/~cfreeman/software/gwas/gtool.html

Haploview http://www.broadinstitute.org/haploview

IMPUTE https://mathgen.stats.ox.ac.uk/impute/impute.html

META http://www.stats.ox.ac.uk/~jsliu/meta.html

METACORE http://www.genego.com/metacore.php

NCBI database Entrez Gene http://www.ncbi.nlm.nih.gov/gene

PLINK http://pngu.mgh.harvard.edu/~purcell/plink/

R http://www.r-project.org/

SNAP http://www.broadinstitute.org/mpg/snap/

SNPTEST http://www.stats.ox.ac.uk/~marchini/software/gwas/snptest_v1.1.5.html

UCSC Genome Browser http://genome.ucsc.edu/

Aaltonen, L., L. Johns, et al. (2007). "Explaining the familial colorectal cancer risk associated with mismatch repair (MMR)-deficient and MMR-stable tumors." Clin Cancer Res **13**(1): 356-61.

Ahmadiyeh, N., M. M. Pomerantz, et al. (2010). "8q24 prostate, breast, and colon cancer risk loci show tissue-specific long-range interaction with MYC." Proc Natl Acad Sci U S A **107**(21): 9742-6.

Ait-Tahar, K., A. P. Liggins, et al. (2009). "Cytolytic T-cell response to the PASD1 cancer testis antigen in patients with diffuse large B-cell lymphoma." Br J Haematol **146**(4): 396-407.

Al-Tassan, N., N. H. Chmiel, et al. (2002). "Inherited variants of MYH associated with somatic G:C-->T:A mutations in colorectal tumors." Nat Genet **30**(2): 227-32.

Altshuler, D., V. J. Pollara, et al. (2000). "An SNP map of the human genome generated by reduced representation shotgun sequencing." Nature **407**(6803): 513-6.

Amundadottir, L. T., P. Sulem, et al. (2006). "A common variant associated with prostate cancer in European and African populations." Nat Genet **38**(6): 652-8.

Anderson, C. A., F. H. Pettersson, et al. (2008). "Evaluating the effects of imputation on the power, coverage, and cost efficiency of genome-wide SNP platforms." Am J Hum Genet **83**(1): 112-9.

Arber, N., C. J. Eagle, et al. (2006). "Celecoxib for the prevention of colorectal adenomatous polyps." N Engl J Med **355**(9): 885-95.

Ardlie, K. G., L. Kruglyak, et al. (2002). "Patterns of linkage disequilibrium in the human genome." Nat Rev Genet **3**(4): 299-309.

Assie, G., T. LaFramboise, et al. (2008). "Frequency of germline genomic homozygosity associated with cancer cases." Jama **299**(12): 1437-45.

Bacolod, M. D., G. S. Schemmann, et al. (2009). "Emerging paradigms in cancer genetics: some important findings from high-density single nucleotide polymorphism array studies." Cancer Res **69**(3): 723-7.

Bacolod, M. D., G. S. Schemmann, et al. (2008). "The signatures of autozygosity among patients with colorectal cancer." Cancer Res **68**(8): 2610-21.

Bercovici, S., C. Meek, et al. (2010). "Estimating genome-wide IBD sharing from SNP data via an efficient hidden Markov model of LD with application to gene mapping." Bioinformatics **26**(12): i175-82.

Bingham, S. A., T. Norat, et al. (2005). "Is the association with fiber from foods in colorectal cancer confounded by folate intake?" Cancer Epidemiol Biomarkers Prev **14**(6): 1552-6.

Bodmer, W. and I. Tomlinson (2010). "Rare genetic variants and the risk of cancer." Curr Opin Genet Dev **20**(3): 262-7.

Bodmer, W. F. (1973). "Genetic factors in Hodgkin's disease: association with a disease-susceptibility locus (DSA) in the HL-A region." Natl Cancer Inst Monogr **36**: 127-34.

Bodmer, W. F., C. J. Bailey, et al. (1987). "Localization of the gene for familial adenomatous polyposis on chromosome 5." Nature **328**(6131): 614-6.

Boehnke, M. (1994). "Limits of resolution of genetic linkage studies: implications for the positional cloning of human disease genes." Am J Hum Genet **55**(2): 379-90.

Boland, C. R., S. N. Thibodeau, et al. (1998). "A National Cancer Institute Workshop on Microsatellite Instability for cancer detection and familial predisposition:

development of international criteria for the determination of microsatellite instability in colorectal cancer." Cancer Res **58**(22): 5248-57.

Bonaiti-Pellie, C. (1999). "Genetic risk factors in colorectal cancer." Eur J Cancer Prev **8 Suppl 1**: S27-32.

Botstein, D. and N. Risch (2003). "Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease." Nat Genet **33 Suppl**: 228-37.

Botstein, D., R. L. White, et al. (1980). "Construction of a genetic linkage map in man using restriction fragment length polymorphisms." Am J Hum Genet **32**(3): 314-31.

Broderick, P., L. Carvajal-Carmona, et al. (2007). "A genome-wide association study shows that common alleles of SMAD7 influence colorectal cancer risk." Nat Genet **39**(11): 1315-7.

Broman, K. W., J. C. Murray, et al. (1998). "Comprehensive human genetic maps: individual and sex-specific variation in recombination." Am J Hum Genet **63**(3): 861-9.

Broman, K. W. and J. L. Weber (1999). "Long homozygous chromosomal segments in reference families from the centre d'Etude du polymorphisme humain." Am J Hum Genet **65**(6): 1493-500.

Bronner, C. E., S. M. Baker, et al. (1994). "Mutation in the DNA mismatch repair gene homologue hMLH1 is associated with hereditary non-polyposis colon cancer." Nature **368**(6468): 258-61.

Brosens, L. A., A. van Hattem, et al. (2007). "Risk of colorectal cancer in juvenile polyposis." Gut **56**(7): 965-7.

Campbell, C. D., E. L. Ogburn, et al. (2005). "Demonstrating stratification in a European American population." Nat Genet **37**(8): 868-72.

Cardon, L. R. and L. J. Palmer (2003). "Population stratification and spurious allelic association." Lancet **361**(9357): 598-604.

Carvajal-Carmona, L. G. (2010). "Challenges in the identification and use of rare disease-associated predisposition variants." Curr Opin Genet Dev **20**(3): 277-81.

Cavenee, W. K., T. P. Dryja, et al. (1983). "Expression of recessive alleles by chromosomal mechanisms in retinoblastoma." Nature **305**(5937): 779-84.

Cazier, J.-B. and I. Tomlinson (2009). "General lessons from large-scale studies to identify human cancer predisposition genes." The Journal of Pathology **220**(2): 255-262.

Chakrabarty, S., V. Radjendirane, et al. (2003). "Extracellular calcium and calcium sensing receptor function in human colon carcinomas: promotion of E-cadherin expression and suppression of beta-catenin/TCF activation." Cancer Res **63**(1): 67-71.

Chen, C. D., M. F. Yen, et al. (2003). "A case-cohort study for the disease natural history of adenoma-carcinoma and de novo carcinoma and surveillance of colon and rectum after polypectomy: implication for efficacy of colonoscopy." Br J Cancer **88**(12): 1866-73.

Cho, E., S. A. Smith-Warner, et al. (2004). "Alcohol intake and colorectal cancer: a pooled analysis of 8 cohort studies." Ann Intern Med **140**(8): 603-13.

Colella, S., C. Yau, et al. (2007). "QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data." Nucleic Acids Res **35**(6): 2013-25.

Collins, F. S., L. D. Brooks, et al. (1998). "A DNA polymorphism discovery resource for research on human genetic variation." Genome Res **8**(12): 1229-31.

Cooper, C. D., A. P. Liggins, et al. (2006). "PASD1, a DLBCL-associated cancer testis antigen and candidate for lymphoma immunotherapy." Leukemia **20**(12): 2172-4.

Cummings, J. H. and S. A. Bingham (1998). "Diet and the prevention of cancer." BMJ **317**(7173): 1636-40.

Curtis, D., A. E. Vine, et al. (2008). "Study of regions of extended homozygosity provides a powerful method to explore haplotype structure of human populations." Ann Hum Genet **72**(Pt 2): 261-78.

Daly, M. J., J. D. Rioux, et al. (2001). "High-resolution haplotype structure in the human genome." Nat Genet **29**(2): 229-32.

Denic, S., C. Frampton, et al. (2007). "Risk of cancer in an inbred population." Cancer Detect Prev **31**(4): 263-9.

Devlin, B. and K. Roeder (1999). "Genomic control for association studies." Biometrics **55**(4): 997-1004.

Di Stefano, L., M. R. Jensen, et al. (2003). "E2F7, a novel E2F featuring DP-independent repression of a subset of E2F-regulated genes." EMBO J **22**(23): 6289-98.

Di, X., H. Matsuzaki, et al. (2005). "Dynamic model based algorithms for screening and genotyping over 100 K SNPs on oligonucleotide microarrays." Bioinformatics **21**(9): 1958-63.

Dong, L. M., J. D. Potter, et al. (2008). "Genetic susceptibility to cancer: the role of polymorphisms in candidate genes." Jama **299**(20): 2423-36.

Dubois, P. C., G. Trynka, et al. (2010). "Multiple common variants for celiac disease influencing immune gene expression." Nat Genet **42**(4): 295-302.

Dukes, C. (1932). "The classification of cancer of the rectum." Journal of Pathology and Bacteriology **35**: 323-32.

Easton, D. F., K. A. Pooley, et al. (2007). "Genome-wide association study identifies novel breast cancer susceptibility loci." Nature **447**(7148): 1087-93.

Enciso-Mora, V., F. J. Hosking, et al. (2010). "Risk of breast and prostate cancer is not associated with increased homozygosity in outbred populations." Eur J Hum Genet **18**(8): 909-14.

Eng, C. and H. Ji (1998). "Molecular classification of the inherited hamartoma polyposis syndromes: clearing the muddied waters." Am J Hum Genet **62**(5): 1020-2.

Eussen, S. J., S. E. Vollset, et al. (2010). "Plasma folate, related genetic variants, and colorectal cancer risk in EPIC." Cancer Epidemiol Biomarkers Prev **19**(5): 1328-40.

Fay, J. C., G. J. Wyckoff, et al. (2001). "Positive and negative selection on the human genome." Genetics **158**(3): 1227-34.

Fearon, E. R. and B. Vogelstein (1990). "A genetic model for colorectal tumorigenesis." Cell **61**(5): 759-67.

Fishel, R., M. K. Lescoe, et al. (1993). "The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer." Cell **75**(5): 1027-38.

Frayling, I. M., N. E. Beck, et al. (1998). "The APC variants I1307K and E1317Q are associated with colorectal tumors, but not always with a family history." Proc Natl Acad Sci U S A **95**(18): 10722-7.

Futreal, P. A., L. Coin, et al. (2004). "A census of human cancer genes." Nat Rev Cancer **4**(3): 177-83.

Gaasenbeek, M., K. Howarth, et al. (2006). "Combined array-comparative genomic hybridization and single-nucleotide polymorphism-loss of heterozygosity analysis reveals complex changes and multiple forms of chromosomal instability in colorectal cancers." Cancer Res **66**(7): 3471-9.

Gammon, A., K. Jasperson, et al. (2009). "Hamartomatous polyposis syndromes." Best Pract Res Clin Gastroenterol **23**(2): 219-31.

Ghoussaini, M., H. Song, et al. (2008). "Multiple loci with different cancer specificities within the 8q24 gene desert." J Natl Cancer Inst **100**(13): 962-6.

Gibson, J., N. E. Morton, et al. (2006). "Extended tracts of homozygosity in outbred human populations." Hum Mol Genet **15**(5): 789-95.

Gudbjartsson, D. F., T. Thorvaldsson, et al. (2005). "Allegro version 2." Nat Genet **37**(10): 1015-6.

Gudmundsson, J., P. Sulem, et al. (2008). "Common sequence variants on 2p15 and Xp11.22 confer susceptibility to prostate cancer." Nat Genet **40**(3): 281-3.

Gunderson, K. L., F. J. Steemers, et al. (2005). "A genome-wide scalable SNP genotyping assay using microarray technology." Nat Genet **37**(5): 549-54.

Gusella, J. F., N. S. Wexler, et al. (1983). "A polymorphic DNA marker genetically linked to Huntington's disease." Nature **306**(5940): 234-8.

Haiman, C. A., L. Le Marchand, et al. (2007). "A common genetic risk factor for colorectal and prostate cancer." Nat Genet **39**(8): 954-6.

Haiman, C. A., N. Patterson, et al. (2007). "Multiple regions within 8q24 independently affect risk for prostate cancer." Nat Genet **39**(5): 638-44.

Hajkova, P., K. Ancelin, et al. (2008). "Chromatin dynamics during epigenetic reprogramming in the mouse germ line." Nature **452**(7189): 877-81.

Hajnoczky, G., E. Davies, et al. (2003). "Calcium signaling and apoptosis." Biochem Biophys Res Commun **304**(3): 445-54.

Half, E. and N. Arber (2009). "Colon cancer: preventive agents and the present status of chemoprevention." Expert Opin Pharmacother **10**(2): 211-9.

Haq, A. I., J. Schneeweiss, et al. (2009). "The Dukes staging system: a cornerstone in the clinical management of colorectal cancer." Lancet Oncol **10**(11): 1128.

Hardy, J. and A. Singleton (2009). "Genomewide association studies and human disease." N Engl J Med **360**(17): 1759-68.

Hein, D. W. (2002). "Molecular genetics and function of NAT1 and NAT2: role in aromatic amine metabolism and carcinogenesis." Mutat Res **506-507**: 65-77.

Hemminki, A., D. Markie, et al. (1998). "A serine/threonine kinase gene defective in Peutz-Jeghers syndrome." Nature **391**(6663): 184-7.

Hemminki, A., I. Tomlinson, et al. (1997). "Localization of a susceptibility locus for Peutz-Jeghers syndrome to 19p using comparative genomic hybridization and targeted linkage analysis." Nat Genet **15**(1): 87-90.

Herrera, L., S. Kakati, et al. (1986). "Gardner syndrome in a man with an interstitial deletion of 5q." Am J Med Genet **25**(3): 473-6.

Hirschhorn, J. N. and M. J. Daly (2005). "Genome-wide association studies for common diseases and complex traits." Nat Rev Genet **6**(2): 95-108.

Holsinger, K. E. and B. S. Weir (2009). "Genetics in geographically structured populations: defining, estimating and interpreting F(ST)." Nat Rev Genet **10**(9): 639-50.

Hosking, F. J., E. Papaemmanuil, et al. (2010). "Genome-wide homozygosity signatures and childhood acute lymphoblastic leukemia risk." Blood **115**(22): 4472-7.

Houlston, R. S., J. Cheadle, et al. (2010). "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 **42**(11): 973-7.

Houlston, R. S. and I. P. Tomlinson (2001). "Polymorphisms and colorectal tumor risk." Gastroenterology **121**(2): 282-301.

Houlston, R. S., E. Webb, et al. (2008). "Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer." Nat Genet **40**(12): 1426-35.

Howarth, K., S. Ranta, et al. (2009). "A mitotic recombination map proximal to the APC locus on chromosome 5q and assessment of influences on colorectal cancer risk." BMC Med Genet **10**: 54.

Howe, J. R., J. L. Bair, et al. (2001). "Germline mutations of the gene encoding bone morphogenetic protein receptor 1A in juvenile polyposis." Nat Genet **28**(2): 184-7.

Howe, J. R., S. Roth, et al. (1998). "Mutations in the SMAD4/DPC4 gene in juvenile polyposis." Science **280**(5366): 1086-8.

Howie, B. N., P. Donnelly, et al. (2009). "A flexible and accurate genotype imputation method for the next generation of genome-wide association studies." PLoS Genet **5**(6): e1000529.

Hugot, J. P., M. Chamaillard, et al. (2001). "Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease." Nature **411**(6837): 599-603.

Ioannidis, J. P., G. Thomas, et al. (2009). "Validating, augmenting and refining genome-wide association signals." Nat Rev Genet **10**(5): 318-29.

Jaeger, E., E. Webb, et al. (2008). "Common genetic variants at the CRAC1 (HMPS) locus on chromosome 15q13.3 influence colorectal cancer risk." Nat Genet **40**(1): 26-8.

Jaeger, E. E., K. L. Woodford-Richens, et al. (2003). "An ancestral Ashkenazi haplotype at the HMPS/CRAC1 locus on 15q13-q14 is associated with hereditary mixed polyposis syndrome." Am J Hum Genet **72**(5): 1261-7.

Janssens, B., S. Goossens, et al. (2001). "alphaT-catenin: a novel tissue-specific beta-catenin-binding protein mediating strong cell-cell adhesion." J Cell Sci **114**(Pt 17): 3177-88.

Jasperson, K. W., T. M. Tuohy, et al. (2010). "Hereditary and familial colon cancer." Gastroenterology **138**(6): 2044-58.

Kastler, S., L. Honold, et al. (2010). "POU5F1P1, a putative cancer susceptibility gene, is overexpressed in prostatic carcinoma." Prostate **70**(6): 666-74.

Kemp, Z., L. Carvajal-Carmona, et al. (2006). "Evidence for a colorectal cancer susceptibility locus on chromosome 3q21-q24 from a high-density SNP genome-wide linkage scan." Hum Mol Genet **15**(19): 2903-10.

Kemp, Z. E., L. G. Carvajal-Carmona, et al. (2006). "Evidence of linkage to chromosome 9q22.33 in colorectal cancer kindreds from the United Kingdom." Cancer Res **66**(10): 5003-6.

Kiemeney, L. A., S. Thorlacius, et al. (2008). "Sequence variant on 8q24 confers susceptibility to urinary bladder cancer." Nat Genet **40**(11): 1307-12.

Knudson, A. G. (2001). "Two genetic hits (more or less) to cancer." Nat Rev Cancer **1**(2): 157-62.

Knudson, A. G., Jr. (1971). "Mutation and cancer: statistical study of retinoblastoma." Proc Natl Acad Sci U S A **68**(4): 820-3.

Kong, A., D. F. Gudbjartsson, et al. (2002). "A high-resolution recombination map of the human genome." Nat Genet **31**(3): 241-7.

Kruglyak, L. (1997). "The use of a genetic map of biallelic markers in linkage studies." Nat Genet **17**(1): 21-4.

Kruglyak, L. and D. A. Nickerson (2001). "Variation is the spice of life." Nat Genet **27**(3): 234-6.

Laken, S. J., G. M. Petersen, et al. (1997). "Familial colorectal cancer in Ashkenazim due to a hypermutable tract in APC." Nat Genet **17**(1): 79-83.

Lamlum, H., M. Ilyas, et al. (1999). "The type of somatic mutation at APC in familial adenomatous polyposis is determined by the site of the germline mutation: a new facet to Knudson's 'two-hit' hypothesis." Nat Med **5**(9): 1071-5.

Lamprecht, S. A. and M. Lipkin (2003). "Chemoprevention of colon cancer by calcium, vitamin D and folate: molecular mechanisms." Nat Rev Cancer **3**(8): 601-14.

Lander, E. S. (1996). "The new genomics: global views of biology." Science **274**(5287): 536-9.

Lander, E. S., L. M. Linton, et al. (2001). "Initial sequencing and analysis of the human genome." Nature **409**(6822): 860-921.

Lander, E. S. and N. J. Schork (1994). "Genetic dissection of complex traits." Science **265**(5181): 2037-48.

Larsson, S. C. and A. Wolk (2006). "Meat consumption and risk of colorectal cancer: a meta-analysis of prospective studies." Int J Cancer **119**(11): 2657-64.

Leach, F. S., N. C. Nicolaides, et al. (1993). "Mutations of a mutS homolog in hereditary nonpolyposis colorectal cancer." Cell **75**(6): 1215-25.

Lebel, R. R. and W. B. Gallagher (1989). "Wisconsin consanguinity studies. II: Familial adenocarcinomatosis." Am J Med Genet **33**(1): 1-6.

Lewontin, R. C. (1964). "The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models." Genetics **49**(1): 49-67.

Li, L. C., R. M. Chui, et al. (2000). "A single nucleotide polymorphism in the E-cadherin gene promoter alters transcriptional activities." Cancer Res **60**(4): 873-6.

Lichtenstein, P., N. V. Holm, et al. (2000). "Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland." N Engl J Med **343**(2): 78-85.

Liu, J. Z., F. Tozzi, et al. (2010). "Meta-analysis and imputation refines the association of 15q25 with smoking quantity." Nat Genet **42**(5): 436-40.

Lockhart-Mummery, P. (1925). "CANCER AND HEREDITY." The Lancet **205**(5296): 427-429.

Lynch, E. D., E. A. Ostermeyer, et al. (1997). "Inherited mutations in PTEN that are associated with breast cancer, cowden disease, and juvenile polyposis." Am J Hum Genet **61**(6): 1254-60.

Major, M. B., N. D. Camp, et al. (2007). "Wilms tumor suppressor WTX negatively regulates WNT/beta-catenin signaling." Science **316**(5827): 1043-6.

Marchini, J. and B. Howie (2010). "Genotype imputation for genome-wide association studies." Nat Rev Genet **11**(7): 499-511.

Marchini, J., B. Howie, et al. (2007). "A new multipoint method for genome-wide association studies by imputation of genotypes." Nat Genet **39**(7): 906-13.

Marsh, D. J., J. B. Kum, et al. (1999). "PTEN mutation spectrum and genotype-phenotype correlations in Bannayan-Riley-Ruvalcaba syndrome suggest a single entity with Cowden syndrome." Hum Mol Genet **8**(8): 1461-72.

Matise, T. C., R. Sachidanandam, et al. (2003). "A 3.9-centimorgan-resolution human single-nucleotide polymorphism linkage map and screening set." Am J Hum Genet **73**(2): 271-84.

McPeek, M. S. (1999). "Optimal allele-sharing statistics for genetic mapping using affected relatives." Genet Epidemiol **16**(3): 225-49.

McQuillan, R., A. L. Leutenegger, et al. (2008). "Runs of homozygosity in European populations." Am J Hum Genet **83**(3): 359-72.

Metzker, M. L. (2010). "Sequencing technologies - the next generation." Nat Rev Genet **11**(1): 31-46.

Miyaki, M., M. Konishi, et al. (1997). "Germline mutation of MSH6 as the cause of hereditary nonpolyposis colorectal cancer." Nat Genet **17**(3): 271-2.

Moghaddam, A. A., M. Woodward, et al. (2007). "Obesity and risk of colorectal cancer: a meta-analysis of 31 studies with 70,000 events." Cancer Epidemiol Biomarkers Prev **16**(12): 2533-47.

Morris, A. P. and L. R. Cardon (2007). Whole Genome Association. Handbook of Statistical Genetics. D. J. Balding, Bishop, M. and Cannings, C. Chichester, UK, John Wiley and Sons. **2:** 1238-1263.

Morris, A. P., E. Zeggini, C. M. Lindgren (2009). "Identification of novel putative rheumatoid arthritis susceptibility genes via analysis of rare variants." *BMC Proc.* **3** (Suppl 7):S131.

Mullikin, J. C., S. E. Hunt, et al. (2000). "An SNP map of human chromosome 22." Nature **407**(6803): 516-20.

Murray, J. C., K. H. Buetow, et al. (1994). "A comprehensive human linkage map with centimorgan density. Cooperative Human Linkage Center (CHLC)." Science **265**(5181): 2049-54.

Nakagawa, H., J. C. Lockman, et al. (2004). "Mismatch repair gene PMS2: disease-causing germline mutations are frequent in patients whose tumors stain negative for PMS2 protein, but paralogous genes obscure mutation detection and interpretation." Cancer Res **64**(14): 4721-7.

Niittymaki, I., E. Kaasinen, et al. (2010). "Low-penetrance susceptibility variants in familial colorectal cancer." Cancer Epidemiol Biomarkers Prev **19**(6): 1478-83.

Norat, T., S. Bingham, et al. (2005). "Meat, fish, and colorectal cancer risk: the European Prospective Investigation into cancer and nutrition." J Natl Cancer Inst **97**(12): 906-16.

Ogura, Y., D. K. Bonen, et al. (2001). "A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease." Nature **411**(6837): 603-6.

Ott, J. (1999). Analysis of Human Genetic Linkage. Baltimore and London, The John Hopkins University Press.

Panagopoulos, I., E. Moller, et al. (2008). "The POU5F1P1 pseudogene encodes a putative protein similar to POU5F1 isoform 1." Oncol Rep **20**(5): 1029-33.

Papadopoulos, N., N. C. Nicolaides, et al. (1994). "Mutation of a mutL homolog in hereditary colon cancer." Science **263**(5153): 1625-9.

Papaemmanuil, E., L. Carvajal-Carmona, et al. (2008). "Deciphering the genetics of hereditary non-syndromic colorectal cancer." Eur J Hum Genet **16**(12): 1477-86.

Park, J. Y., P. N. Mitrou, et al. (2010). "Lifestyle factors and p53 mutation patterns in colorectal cancer patients in the EPIC-Norfolk study." Mutagenesis.

Parkin, D. M., F. Bray, et al. (2005). "Global cancer statistics, 2002." CA Cancer J Clin **55**(2): 74-108.

Pei, Y. F., J. Li, et al. (2008). "Analyses and comparison of accuracy of different genotype imputation methods." PLoS One **3**(10): e3551.

Penrose, L. S. (1955). "Evidence of heterosis in man." Proc R Soc Lond B Biol Sci **144**(915): 203-13.

Pettiti, D. (1994). Meta-analysis Decision Analysis and Cost-Effectiveness Analysis. New York, University Press

Pharoah, P. D., A. M. Dunning, et al. (2004). "Association studies for finding cancer-susceptibility genetic variants." Nat Rev Cancer **4**(11): 850-60.

Picelli, S., J. Vandrovcova, et al. (2008). "Genome-wide linkage scan for colorectal cancer susceptibility genes supports linkage to chromosome 3q." BMC Cancer **8**: 87.

Pittman, A. M., E. Webb, et al. (2008). "Refinement of the basis and impact of common 11q23.1 variation to the risk of developing colorectal cancer." Hum Mol Genet **17**(23): 3720-7.

Porter, T. R., F. M. Richards, et al. (2002). "Contribution of cyclin d1 (CCND1) and E-cadherin (CDH1) polymorphisms to familial and sporadic colorectal cancer." Oncogene **21**(12): 1928-33.

Price, A. L., N. J. Patterson, et al. (2006). "Principal components analysis corrects for stratification in genome-wide association studies." Nat Genet **38**(8): 904-9.

Price, A. L., M. E. Weale, et al. (2008). "Long-range LD can confound genome scans in admixed populations." Am J Hum Genet **83**(1): 132-5; author reply 135-9.

Pritchard, J. K. (2001). "Are rare variants responsible for susceptibility to complex diseases?" Am J Hum Genet **69**(1): 124-37.

Pritchard, J. K., M. Stephens, et al. (2000). "Inference of population structure using multilocus genotype data." Genetics **155**(2): 945-59.

Purcell, S., B. Neale, et al. (2007). "PLINK: a tool set for whole-genome association and population-based linkage analyses." Am J Hum Genet **81**(3): 559-75.

Rahman, N., S. Seal, et al. (2007). "PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene." Nat Genet **39**(2): 165-7.

Reich, D. E., S. B. Gabriel, et al. (2003). "Quality and completeness of SNP databases." Nat Genet **33**(4): 457-8.

Resnick, K. E., H. Hampel, et al. (2009). "Current and emerging trends in Lynch syndrome identification in women with endometrial cancer." Gynecol Oncol **114**(1): 128-34.

Rioux, J. D., M. J. Daly, et al. (2001). "Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease." Nat Genet **29**(2): 223-8.

Risch, N. and K. Merikangas (1996). "The future of genetic studies of complex human diseases." Science **273**(5281): 1516-7.

Roberts-Thomson, I. C., P. Ryan, et al. (1996). "Diet, acetylator phenotype, and risk of colorectal neoplasia." Lancet **347**(9012): 1372-4.

Rudan, I., D. Rudan, et al. (2003). "Inbreeding and risk of late onset complex disease." J Med Genet **40**(12): 925-32.

Sachidanandam, R., D. Weissman, et al. (2001). "A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms." Nature **409**(6822): 928-33.

Sahota, S. S., C. M. Goonewardena, et al. (2006). "PASD1 is a potential multiple myeloma-associated antigen." Blood **108**(12): 3953-5.

Satagopan, J. M., D. A. Verbel, et al. (2002). "Two-stage designs for gene-disease association studies." Biometrics **58**(1): 163-70.

Sieber, O. M., L. Lipton, et al. (2003). "Multiple colorectal adenomas, classic adenomatous polyposis, and germ-line mutations in MYH." N Engl J Med **348**(9): 791-9.

Sieber, O. M., S. Segditsas, et al. (2006). "Disease severity and genetic pathways in attenuated familial adenomatous polyposis vary greatly but depend on the site of the germline mutation." Gut **55**(10): 1440-8.

Skol, A. D., L. J. Scott, et al. (2006). "Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies." Nat Genet **38**(2): 209-13.

Slager, S. L., J. Huang, et al. (2000). "Effect of allelic heterogeneity on the power of the transmission disequilibrium test." Genet Epidemiol **18**(2): 143-56.

Sotelo, J., D. Esposito, et al. (2010). "Long-range enhancers on 8q24 regulate c-Myc." Proc Natl Acad Sci U S A **107**(7): 3001-5.

Spain, S. L., J. B. Cazier, et al. (2009). "Colorectal cancer risk is not associated with increased levels of homozygosity in a population from the United Kingdom." Cancer Res **69**(18): 7422-9.

Spencer, C. C., Z. Su, et al. (2009). "Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip." PLoS Genet **5**(5): e1000477.

Stoetzel, C., V. Laurier, et al. (2006). "BBS10 encodes a vertebrate-specific chaperonin-like protein and is a major BBS locus." Nat Genet **38**(5): 521-4.

Stoll, M., B. Corneliussen, et al. (2004). "Genetic variation in DLG5 is associated with inflammatory bowel disease." Nat Genet **36**(5): 476-80.

Syngal, S., E. A. Fox, et al. (2000). "Sensitivity and specificity of clinical criteria for hereditary non-polyposis colorectal cancer associated mutations in MSH2 and MLH1." J Med Genet **37**(9): 641-5.

Taioli, E., M. A. Garza, et al. (2009). "Meta- and pooled analyses of the methylenetetrahydrofolate reductase (MTHFR) C677T polymorphism and colorectal cancer: a HuGE-GSEC review." Am J Epidemiol **170**(10): 1207-21.

Takeichi, M. (1991). "Cadherin cell adhesion receptors as a morphogenetic regulator." Science **251**(5000): 1451-5.

Tenesa, A., S. M. Farrington, et al. (2008). "Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21." Nat Genet **40**(5): 631-7.

The International HapMap Consortium (2003). "The International HapMap Project." Nature **426**(6968): 789-796.

The International HapMap Consortium (2005). "A haplotype map of the human genome." Nature **437**(7063): 1299-320.

The International HapMap Consortium, K. A. Frazer, et al. (2007). "A second generation human haplotype map of over 3.1 million SNPs." Nature **449**(7164): 851-61.

Thompson, E., C. J. Meldrum, et al. (2004). "Hereditary non-polyposis colorectal cancer and the role of hPMS2 and hEXO1 mutations." Clin Genet **65**(3): 215-25.

Tomlinson, I., N. Rahman, et al. (1999). "Inherited susceptibility to colorectal adenomas and carcinomas: evidence for a new predisposition gene on 15q14-q22." Gastroenterology **116**(4): 789-95.

Tomlinson, I., E. Webb, et al. (2007). "A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21." Nat Genet **39**(8): 984-8.

Tomlinson, I. P., M. Dunlop, et al. (2010). "COGENT (COlorectal cancer GENeTics): an international consortium to study the role of polymorphic variation on the risk of colorectal cancer." Br J Cancer **102**(2): 447-54.

Tomlinson, I. P., E. Webb, et al. (2008). "A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3." Nat Genet **40**(5): 623-30.

Valerie, K. and L. F. Povirk (2003). "Regulation and mechanisms of mammalian double-strand break repair." Oncogene **22**(37): 5792-812.

Venter, J. C., M. D. Adams, et al. (2001). "The sequence of the human genome." Science **291**(5507): 1304-51.

Verhage, B. A., K. van Houwelingen, et al. (2002). "Single-nucleotide polymorphism in the E-cadherin gene promoter modifies the risk of prostate cancer." Int J Cancer **100**(6): 683-5.

Visscher, P. M., W. G. Hill, et al. (2008). "Heritability in the genomics era--concepts and misconceptions." Nat Rev Genet **9**(4): 255-66.

Vogelstein, B., E. R. Fearon, et al. (1988). "Genetic alterations during colorectal-tumor development." N Engl J Med **319**(9): 525-32.

Wang, S., C. Haynes, et al. (2009). "Genome-wide autozygosity mapping in human populations." Genet Epidemiol **33**(2): 172-80.

Watson, J. D., T. A. Baker, et al. (2004). Molecular Biology of the Gene. San Francisco, Benjamin Cummings.

Wellcome Trust Case Control Consortium (2007). "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls." Nature **447**(7145): 661-78.

Whitelaw, S. C., V. A. Murday, et al. (1997). "Clinical and molecular features of the hereditary mixed polyposis syndrome." Gastroenterology **112**(2): 327-34.

Wiesner, G. L., D. Daley, et al. (2003). "A subset of familial colorectal neoplasia kindreds linked to chromosome 9q22.2-31.2." Proc Natl Acad Sci U S A **100**(22): 12961-5.

Wright, S. (1922). "Coefficients of Inbreeding and Relationship." The American Naturalist **56**(645): 330.

Yang, Q., M. J. Khoury, et al. (2005). "How many genes underlie the occurrence of common complex diseases in the population?" Int J Epidemiol **34**(5): 1129-37.

Yaspan, B. L., K. M. McReynolds, et al. (2008). "A haplotype at chromosome Xq27.2 confers susceptibility to prostate cancer." Hum Genet **123**(4): 379-86.

Yeager, M., N. Orr, et al. (2007). "Genome-wide association study of prostate cancer identifies a second risk locus at 8q24." Nat Genet **39**(5): 645-9.

Zanke, B. W., C. M. Greenwood, et al. (2007). "Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24." Nat Genet **39**(8): 989-94.

Zeggini, E., L. J. Scott, et al. (2008). "Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes." Nat Genet **40**(5): 638-45.

Zollner, S. and J. K. Pritchard (2005). "Coalescent-based association mapping and fine mapping of complex trait loci." Genetics **169**(2): 1071-92.

Chapter 9. Appendix

9.1 Summary description of the datasets included in the GWA study

9.1.1 England phase one (EngP1)

The EngP1 cohorts used in phase one of the association analysis consisted of 930 cases (45% male and 55% female) collected through the Colorectal Tumour Gene Identification (CORGI) consortium (Tomlinson *et al.* 2007). Overall the cases included 620 CRC's and 310 high risk adenomas. All had at least one first degree relative affected with CRC and met at least one of the following criteria:

- CRC at age 75 years or younger
- A colorectal adenoma at age 45 years or younger
- Three or more colorectal adenomas at age 75 years or younger
- One large colorectal adenoma greater than 1cm in diameter at age 75 or younger
- One tubulovillous or severely dysplastic adenoma at age 75 years or younger

The 965 controls for this cohort (45% males and 55% females) were also collected as part of the CORGI consortium and were generally the spouses of the cases that were unaffected and had no known family history of CRC. The known predisposition conditions for polyposis syndromes, APC, HNPCC and MYH were excluded from this study. All samples were of white UK origin, which was self-assessed by questionnaire . These samples were genotyped on the Illumina Hap550 SNP array.

9.1.2 Scotland phase one (ScotP1)

The Scottish phase one samples consisted of 1,012 CRC cases (518 males and 494 females) selected for early age of onset and 1,012 cancer-free population controls (518 males and 494 females) that were age (±5 years), gender and area of residence matched to the cases. These samples were genotyped on the Illumina Hap550 SNP array.

9.1.3 England phase two (EngP2)

The cases were 2,873 CRC patients (1,199 males and 1,674 females) recruited through The National Study of Colorectal Cancer Genetics (NSCCG) and the Royal Marsden Hospital NHS trust and Institute of Cancer Research Family History and DNA Registry. The 2,871 controls (1,164 males and 1,707 females) were made up of unrelated individuals collected as part of NSCCG, the Genetic Lung Cancer Predisposition Study and the Institute of Cancer Research/ Royal Marsden NHS Trust Family history and DNA Registry. All cases and controls are UK Caucasian and with similar demography in terms of place of residence.

9.1.4 Scotland phase two (ScotP2)

ScotP2 is comprised of 2,057 CRC cases (1,249 males and 808 females) aged less than 80 years at the time of diagnosis and 2,111 population controls (1,257 males and 854 females) collected as part of an independent CRC incidence study and matched by age, gender and place of residence.

9.1.5 VQ58

The VQ58 cohort consists of 1,432 CRC cases, which were recruited through two clinical trials of adjuvant therapy (www.octo-oxford.org.uk). These included 929 cases from VICTOR, a phase III randomised double blind placebo study of the drug rofecoxib (VIOXX), all with Dukes B or C CRC and 503 cases from the QUASAR2 trial, which compared chemotherapy using capecitabine against capecitabine and Avastin® (bevacizumab). These samples were genotyped in house using the Illumina Hap300/317 SNP arrays. The controls were 2,697 population controls from the publicly available WTCCC2 1958 birth cohort, which were genotyped using the Illumina 1M SNP array (http://www.wtccc.org.uk/ccc2/wtccc2_studies.shtml).

9.1.6 CFR

The Cancer Family Registry (CFR) cohort consists of 1,186 CRC cases and 998 controls recruited from three centres: Toronto, Melbourne and Seattle, and genotyped using the Illumina 1M SNP arrays.

9.1.7 Australia

The dataset consists of 360 cases and 1,870 controls. The cases were recruited from Melbourne as part of the Ludwig Colon Cancer Initiative and the controls from this dataset were recruited from Brisbane as part of the Queensland Institute of Medical Research (QIMR) studies.

9.1.8 The replication cohorts

The samples used for the replication phase were collected as part of the colorectal cancer genetics consortium (COGENT)(Tomlinson *et al.* 2010). The number of cohorts included in the replication phase has grown throughout the study and, therefore, different cohorts were included in each stage of the analysis. The details are given in the results section.

COIN/NBS – The cases consisted of 2,182 samples recruited through the COIN and COIN-B clinical trials of metastatic CRC. The controls consisted of 2,501 samples were from the publicly available UK National Blood service (NBS) population controls.

NSCCG post 2005 (EngP3) consists of 3,286 cases (2,158 males and 1,128 females) and 3,017 controls (1,212 males and 1,805 males) that were recruited after 2005 as part of NSCCG.

CORGI2bcd consisted of 588 CRC cases collected as part of the CORGI consortium post 2005 and 1,092 cancer-free population, spouse or European Collection of Cell Cultures (ECACC) controls. CORGI2bcd and EngP4, below, overlap with respect to samples; EngP4 was used initially, but was later split in two when VICTOR was genotyped on the Hap370 arrays and combined with QUASAR2 and the 1958 birth cohort controls.

EngP4 consisted of 182 CRC cases from CORGI2bcd and 888 VICTOR CRC cases. Controls consisted of 100 European Collection of Cell Cultures (ECACC) controls and 315 unaffected controls collected through CORGI.

Cambridge (SEARCH) consisted of 2,222 CRC cases (1,278 males and 944 females) and 2,262 controls (949 males, 1,313 females) ascertained through the Studies of Epidemiology and Risk Factors in Cancer Heredity (SEARCH) study based in Cambridge. Controls were sex and age matched to the cases.

Finland Colorectal Cancer Predisposition Study (FCCPS) consisted of 962 CRC cases and 846 controls that were randomly selected from Finnish blood donors .

DACHS consisted of 1,373 CRC cases (790 males and 583 females) and 1,480 controls (719 males and 761 females) recruited as part of a case controls study into the incidence of CRC in the Rhine-Nectar-Odenwald region near Heidelberg between 2003 and 2006.

Kiel consisted of 2,169 CRC cases (1,089 males and 1,080 females) and 2,145 controls, which were collected as part of the SHIP and PopGen population biobank projects based in Kiel and Greifswald, Germany.

Canada consisted of 1,175 CRC cases (503 males and 672 females) and 1,184 controls (667 males and 517 females) collected through the Ontario Familial Colorectal Cancer Registry.

The following cohorts, along with the samples from ScotP1 and ScotP2 were utilised in the replication phase of the combined analysis of EngP1 and EngP2 (Tomlinson *et al.* 2008).

POPGENSHIP consisted of 2,569 cases affected with CRC (1,382 males and 1,187 females) and 2,699 controls (1,296 males and 1,395 females), collected as part of two population based biobank projects, PopGen and SHIP, in Kiel and Greifswald, Germany. The samples in this dataset are included in the Kiel cohort described above.

DFCCS consisted of 783 cases with a family history of disease (370 males and 413 females) and 664 controls (251 males and 413 females) recruited from a genetic reference centre in Leiden, The Netherlands.

MCCS consisted of 515 cases affected with CRC (270 males and 245 females) and 709 controls (352 males and 357 females), selected randomly from a cohort ascertained in Melbourne, Australia as part of the Melbourne Collaborative Cohort Study.

EPICOLON consisted of 515 cases affected with CRC (305 males and 210 females) and 515 controls (290 males and 225 controls) collected from Barcelona, Spain as part of the nationwide EPICOLON project, which compiled epidemiological and clinical information on individuals with HNPCC and other familial cancer conditions.

9.2 Scripts for meta-analysing GWA studies

I formatted the association analysis results correctly using an awk script, as in the following example:

getgrp_allelicgetcounts2.awk_test

```
awk '
BEGIN{  while(getline<"P1_cleaned_170410_m.aff.counts"){
AaC[$2]=(2*$5)+$6
BaC[$2]=(2*$7)+$6
taC[$2]=2*($5+$6+$7)
chraC[$2]=$1
A1aC[$2]=$3
A2aC[$2]=$4
}
while(getline<"P1_cleaned_170410_m.ctrl.counts"){
AcC[$2]=(2*$5)+$6
BcC[$2]=(2*$7)+$6
tcC[$2]=2*($5+$6+$7)
chrcC[$2]=$1
A1cC[$2]=$3
A2cC[$2]=$4
}
while(getline<"VQ58v3clean_170409_m.aff.counts"){
AaV[$2]=(2*$5)+$6
BaV[$2]=(2*$7)+$6
taV[$2]=2*($5+$6+$7)
chraV[$2]=$1
A1aV[$2]=$3
A2aV[$2]=$4
}
while(getline<"VQ58v3clean_170409_m.ctrl.counts"){
AcV[$2]=(2*$5)+$6
BcV[$2]=(2*$7)+$6
tcV[$2]=2*($5+$6+$7)
chrcV[$2]=$1
A1cV[$2]=$3
A2cV[$2]=$4
}
print "done reading counts" > "/dev/stderr"
}
(NR>1){
snp=$2
chr=$1
loc=$3
if (A1aV[snp]!=A1cC[snp]){
aaV=BaV[snp]
      baV=AaV[snp]
      a1aV=A2aV[snp]
      a2aV=A1aV[snp]
      acV=BcV[snp]
      bcV=AcV[snp]
      a1cV=A2cV[snp]
      a2cV=A1cV[snp]
   }else{
    aaV=AaV[snp]
    baV=BaV[snp]
    a1aV=A1aV[snp]
    a2aV=A2aV[snp]
    acV=AcV[snp]
    bcV=BcV[snp]
    a1cV=A1cV[snp]
    a2cV=A2cV[snp]
   }
  sumA=AaC[snp]+AcC[snp]+aaV+acV
  sumB= BaC[snp]+BcC[snp]+baV+bcV
  if (sumA<sumB){
    maC=AaC[snp]
    mcC=AcC[snp]
    maV=aaV
    mcV=acV
    m1aV=a1aV
    m1cV=a1cV
    m1aC=A1aC[snp]
    m1cC=A1cC[snp]
  }else{
    maC=BaC[snp]
    mcC=BcC[snp]
    maV=baV
    mcV=bcV
    m1aV=a2aV
    m1cV=a2cV
    m1aC=A2aC[snp]
    m1cC=A2cC[snp]
  }
  if (maC*mcC*maV*mcV !=0)
     print snp, chr, loc, m1aC, maC, taC[snp], m1cC, mcC, tcC[snp], m1aV, maV, taV[snp],
m1cV, mcV,  vtcV[snp]
}'  P1_cleaned_170410_a.assoc
```

The output from the script above was then used as the input for the meta-analysis R script below. N is the number of SNPs.

```
library(meta)

y<-read.table("p1p2md1md2vq58_110510.metain",sep="
",header=F,colClasses=c("character","numeric","numeric","character","numeric","numeric","
character","numeric","numeric","character","numeric","numeric","character","numeric","nu
meric","character","numeric","numeric","character","numeric","numeric","character","num
eric","numeric","character","numeric","numeric","character","numeric","numeric","characte
r","numeric","numeric"))

write(paste("SNP","Chr","Location","Fixed_P","Fixed_OR","Random_P","Random_OR",sep="
\t"),file="p1p2md1md2vq58_110510.metares")

z<-matrix(rep(1,n*7),n,7)

 for (i in 1:n) {
   met<-
metabin(c(y[i,5],y[i,11],y[i,17],y[i,23],y[i,29],y[i,35]),c(y[i,6],y[i,12],y[i,18],y[i,24],y[i,30],y[i,36]
),c(y[i,8],y[i,14],y[i,20],y[i,26],y[i,32],y[i,38]),c(y[i,9],y[i,15],y[i,21],y[i,27],y[i,33],y[i,39]),sm="
OR")

   z[i,1]=y[i,1]
   z[i,2]=y[i,2]
   z[i,3]=y[i,3]
   if (met$TE.fixed<0) {
     z[i,4]=2*pnorm(met$TE.fixed/met$seTE.fixed)
   }
   if (met$TE.fixed>0) {
     z[i,4]=2*pnorm(-met$TE.fixed/met$seTE.fixed)
   }
   if (met$TE.random<0) {
     z[i,6]=2*pnorm(met$TE.random/met$seTE.random)
   }
   if (met$TE.random>0) {

   z[i,6]=2*pnorm(-met$TE.random/met$seTE.random)
   }
   z[i,5]=exp(met$TE.fixed)
   z[i,7]=exp(met$TE.random)
 }
   print(z[,1])

write.table(z,file="p1p2md1md2vq58_110510.metares",
quote=F,sep="\t",append=T,row.names=F,col.names=FALSE)
```

9.3 Imputation scripts

9.3.1 Convert the SNP positions to match the reference panel:

change_locs.awk

```
#!/usr/bin/awk -f
BEGIN{
  while(getline<"../corgi_550/HumanHap550v3_A.map"){
  loc36[$1]=$2
  }
}
{
loc1=$3
snp=$2
if(loc1!=loc36[snp]){$3=loc36[snp]}
print $0
}
```

This script is run over all chromosomes using *change_locs.sh*

```
for i in `seq 1 22` ; do
 ./change_locs.awk  ../VQv2c_chr${i}.gen | sort –gk3 > VQv2c_chr${i}.hap36.gen
done
```

9.3.2 Create the command file for each 5Mb segment

The following script, *make_multiimpute_com2.awk,* was used to separate each chromosome into 7Mb segments, created a command file for each segment. This script is for IMPUTEv1. I then used the script *run_impute_com.sh* to run IMPUTE across all segments for each chromosome in a loop.

make_multiimpute_com2.awk

```
#!/usr/bin
#to use do ./script.awk chr.length
#the following function performs if x<y return x, otherwise return y

function min(x,y){return x < y ? x:y}
BEGIN{
    L=7000000
    h=0
    i=0
}
{
   while(h<=$2){
   i=i+1
   mystart=h
   myend=min(h+L,$2)
   print "cd /farm/home/spain01/projects/GWA/impute; ./impute
 -h b36_files/hapmap_r24_b36_fwd.consensus.qc.poly."$1"_ceu.phased
 -l b36_files/"$1".ceu.r24.legend
 -m b36_files/genetic_map_"$1"_CEU_b36.txt
 -fix_strand
 -g hapmap_impute1/VQv2c_"$1".hap36.gen
 -Ne 11418 -o hapmap_impute1/"$1"/VQv2c_"$1"_"i".imputed
 -i hapmap_impute1/"$1"/VQv2c_"$1"_"i".info
 -r hapmap_impute1/"$1"/"$1"_"i".summary -int", mystart, myend   > ""$1"/"$1"_"i".com"
h=h+L
   }
}
```

The script is run across chromosomes using *run_impute_com.sh:*

```
#!/bin/sh
for c in `seq 1 22` ; do
./make_multiimpute_com2.awk chr_b36length/chr${c}_length.b36
done
```

A list of all the command files was created for each chromosome called file.list and the actual imputation from the command files was then performed by running the following script:

```
run_impute.multi.com

for i in ‘seq 1 22’; do
cd chr${i}
 for i in `cat file.list` ; do
  sh ${i}
 done
done
```

9.3.3 SNPTEST scripts

```
make_multi.snptest.sh
cd /farm/home/spain01/projects/GWA/impute/

for i in `seq 1 22` ; do

awk -vc="${i}" '
   {print "cd /farm/home/spain01/projects/GWA/impute/ ; ./snptest
-cases hapmap_impute1/chr"c"/VQv2_550."$1".imputed /VQv2/VQv2c.aff.sample
-controls WTCCC2/Hap317/chr"c"/BC58cleaner_317."$1".imputed
WTCCC2/Hap317/BC58_1Mclean.ctrl.sample
-o hapmap_impute1/snptest_hapmap317v2/chr"c"/VQ58v3_550_"$1".snptest.res
-exclude_samples hapmap_impute1/snptest_hapmap317v2/exclusion.sample.list
-frequentist 1 2 -proper -hwe -bayesian 1 2 -nsamp 250"}' chr${i}/imputed.file.id >
"hapmap_impute1/snptest_hapmap317v2/chr${i}/snptest_chr${i}_v2.com"
done
```

This command was then performed across all imputed segments using the following script:

```
run_snptest.com
for i in `seq 2 22` ; do
 cd chr${i}
 qsub -l nodes=1:g6blade48 ./snptest_chr${i}_v2.com
 cd ../
done
```

9.4 Analysis of runs of homozygosity scripts

9.4.1 Analysis of homozygosity by SNP

The analysis of homozygosity by SNP was performed by chi square test using the R script below.

```
chisq_multi.R
mytest=function(x){ # Use Fisher instead of Chisquare if one cell is lower or equal to 5
  mx<-matrix(round(x),nr=2,byrow=TRUE)
  thistest<-TRUE %in% names(table(x<6))
  p<-ifelse(thistest, fisher.test(mx)$p.value,chisq.test(mx)$p.value)
  return(p)
}
aff_counts<-read.table("corgi.aff.counts", header=F)
ctrl_counts<-read.table("corgi.ctrl.counts", header=F)
write(paste("SNP","chr","chisq_pval","homo_OR","AABBaff","AB_aff","AABBctrl","AB_ctrl",s
ep="\t"),file="corgi550_homozpval_040409.dat")
for (i in 1:485179){
      SNP=ctrl_counts[i,2]
      chr=ctrl_counts[i,1]
      AABBaff=(aff_counts[i,5]+aff_counts[i,7])
      ABaff=aff_counts[i,6]
      AABBctrl=(ctrl_counts[i,5]+ctrl_counts[i,7])
      ABctrl=ctrl_counts[i,6]
      m=mytest(c(AABBaff,ABaff,AABBctrl,ABctrl))
      o=((AABBaff/ABaff)/(AABBctrl/ABctrl))

write(paste(SNP,chr,m,o,AABBaff,ABaff,AABBctrl,ABctrl,sep="\t"),file="corgi550_homozpval
_040409.dat",append=T)
      i=i+1
}
```

9.4.2 Meta-analysis of homozygosity association results

Meta-analysis was performed in R using the following script. The input file has one row per SNP consisting of SNP, chromosome, position, homozygote count in cases, total genotype count in cases, homozygote count in controls, and total genotype count in controls.

2grpmeta.R

```
library(meta)

y<-read.table("p1VQ58_homoz_metain4409",sep="
",header=F,colClasses=c("character","character","numeric","numeric","numeric","numeric",
"numeric","numeric","numeric","numeric","numeric"))

z<-matrix(rep(1,290010*7),290010,7)
 for (i in 1:290010) {
   met<-metabin(c(y[i,4],y[i,8]),c(y[i,5],y[i,9]),c(y[i,6],y[i,10]),c(y[i,7],y[i,11]),sm="OR")
   z[i,1]=y[i,1]
   z[i,2]=y[i,2]
   z[i,3]=y[i,3]
   if (met$TE.fixed<0) {
     z[i,4]=2*pnorm(met$TE.fixed/met$seTE.fixed)
   }
   if (met$TE.fixed>0) {
     z[i,4]=2*pnorm(-met$TE.fixed/met$seTE.fixed)
   }
   if (met$TE.random<0) {
     z[i,6]=2*pnorm(met$TE.random/met$seTE.random)
   }
   if (met$TE.random>0) {
     z[i,6]=2*pnorm(-met$TE.random/met$seTE.random)
   }
   z[i,5]=exp(met$TE.fixed)
   z[i,7]=exp(met$TE.random)
 }
   print(z[,1])
write.table(z,file='p1VQ58_300snps_homozmeta060409',quote=F,sep=" “,
row.names=F,col.names=FALSE)
```

9.4.3 Analysis of recurrent ROH regions

The following R script was used for this analysis.

ROH_Pvalue.R

```
mytest=function(x){ # Use Fisher instead of Chi square if one cell is lower or equal to 5
 mx<-matrix(round(x),nr=2,byrow=TRUE)
 thistest<-TRUE %in% names(table(x<6))
 p<-ifelse(thistest, fisher.test(mx)$p.value,chisq.test(mx)$p.value)
 return(p)
}
homo_counts<-read.table("corgi550_hwmaf4_1000_hom.CONvert", header=T)
write(paste("homo_region","chr","start_pos","end_pos","size(kb)","chisq_pval","homo_OR"
,"homoaff","nonhom_aff","homoctrl","nonhom_ctrl",sep="\t"),file="corgi550_1000_pval4_
CON_280409.dat")
for (i in 1:6405){
    homo_region=homo_counts[i,1]
    chr=homo_counts[i,2]
    start_pos=homo_counts[i,3]
    end_pos=homo_counts[i,4]
    size=homo_counts[i,5]
    homoaff=homo_counts[i,6]
    otheraff=(homo_counts[i,7]-homo_counts[i,6])
    homoctrl=homo_counts[i,8]
    otherctrl=(homo_counts[i,9]-homo_counts[i,8])

    m=mytest(c(homoaff,otheraff,homoctrl,otherctrl))
    o=((homoaff/otheraff)/(homoctrl/otherctrl))

write(paste(homo_region,chr,start_pos,end_pos,size,m,o,homoaff,otheraff,homoctrl,otherc
trl,sep="\t"),file="corgi550_1000_pval4_CON_280409.dat",append=T)
    i=i+1
}
```

9.5 The common pathways of genes tagged by associated SNPs

These figures relate to Section 3.8 and illustrate the 3 most gene list enriched pathways .

Figure 9.1 Cadherin-mediated cell adhesion pathway

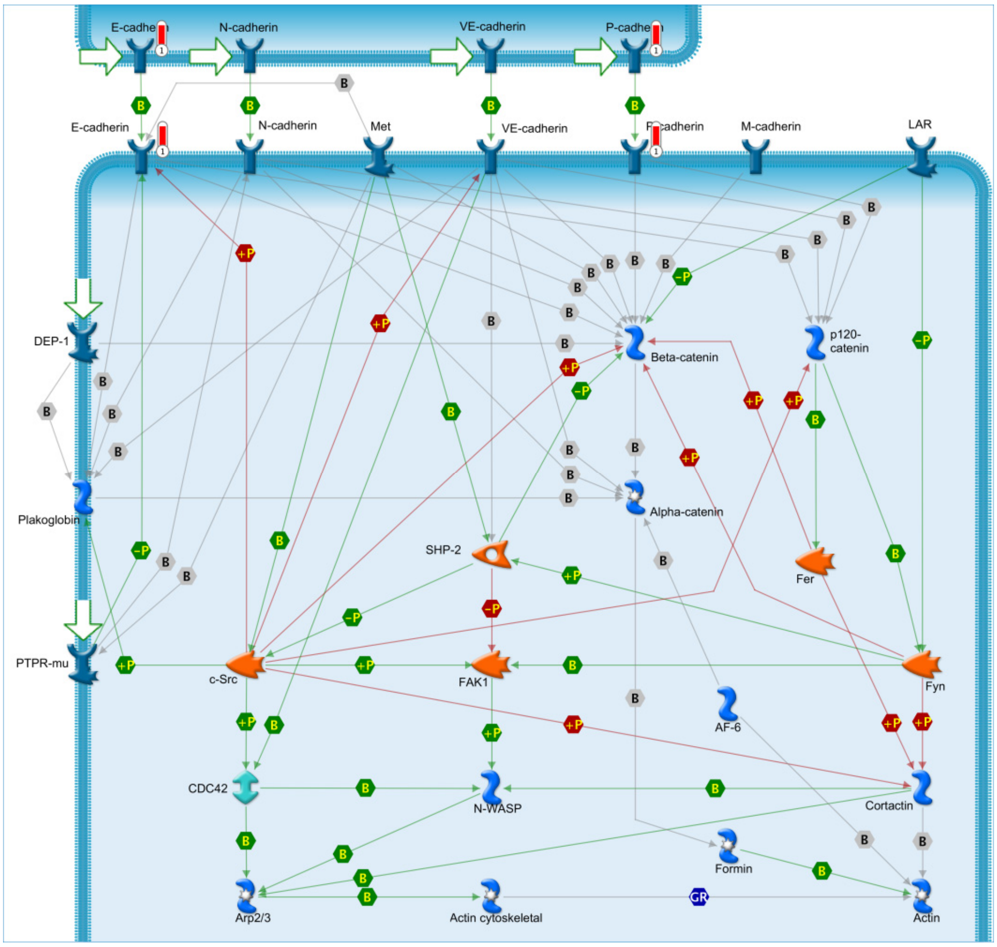


Figure 9.2 Wnt signalling pathway (part 2)

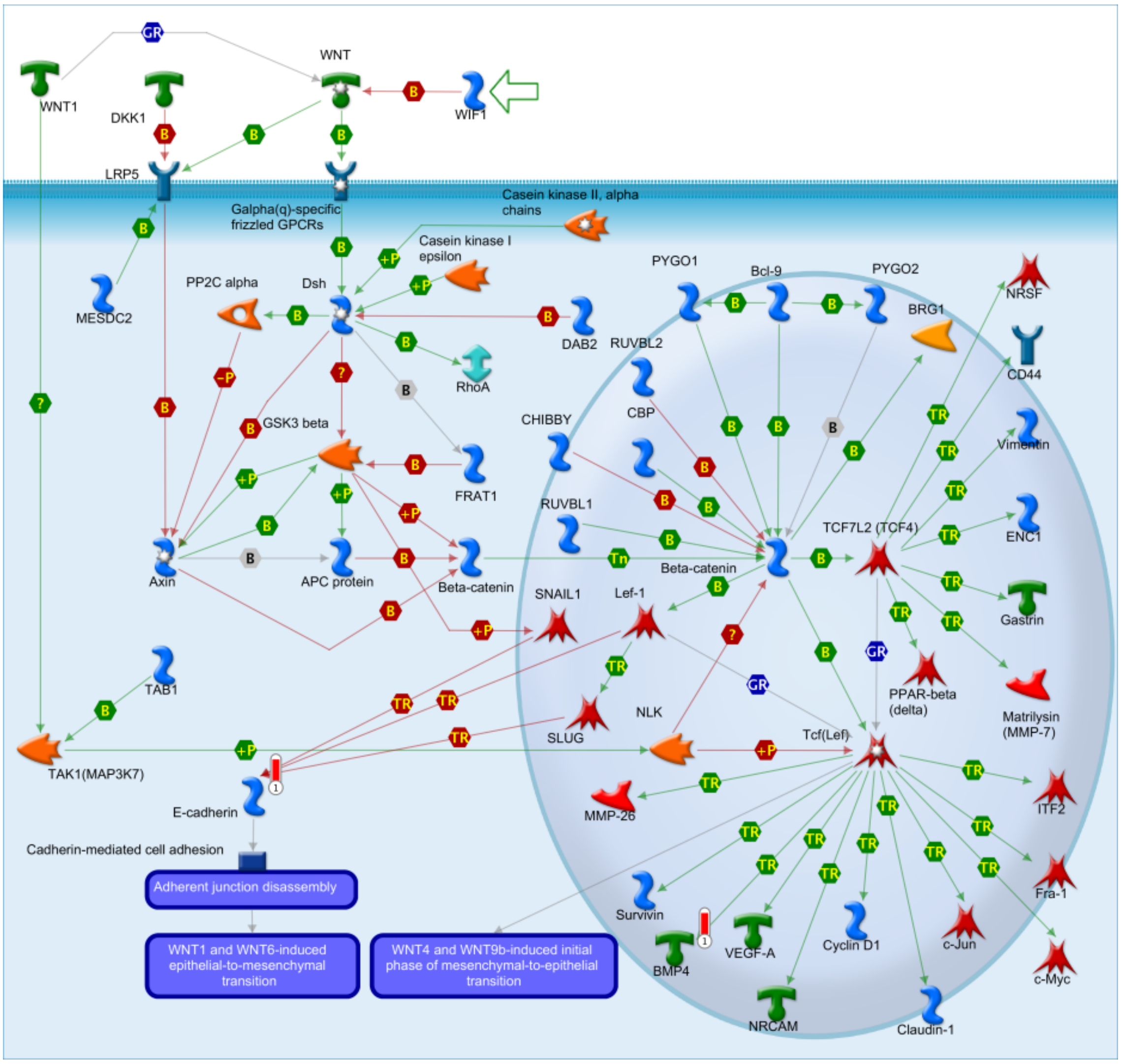
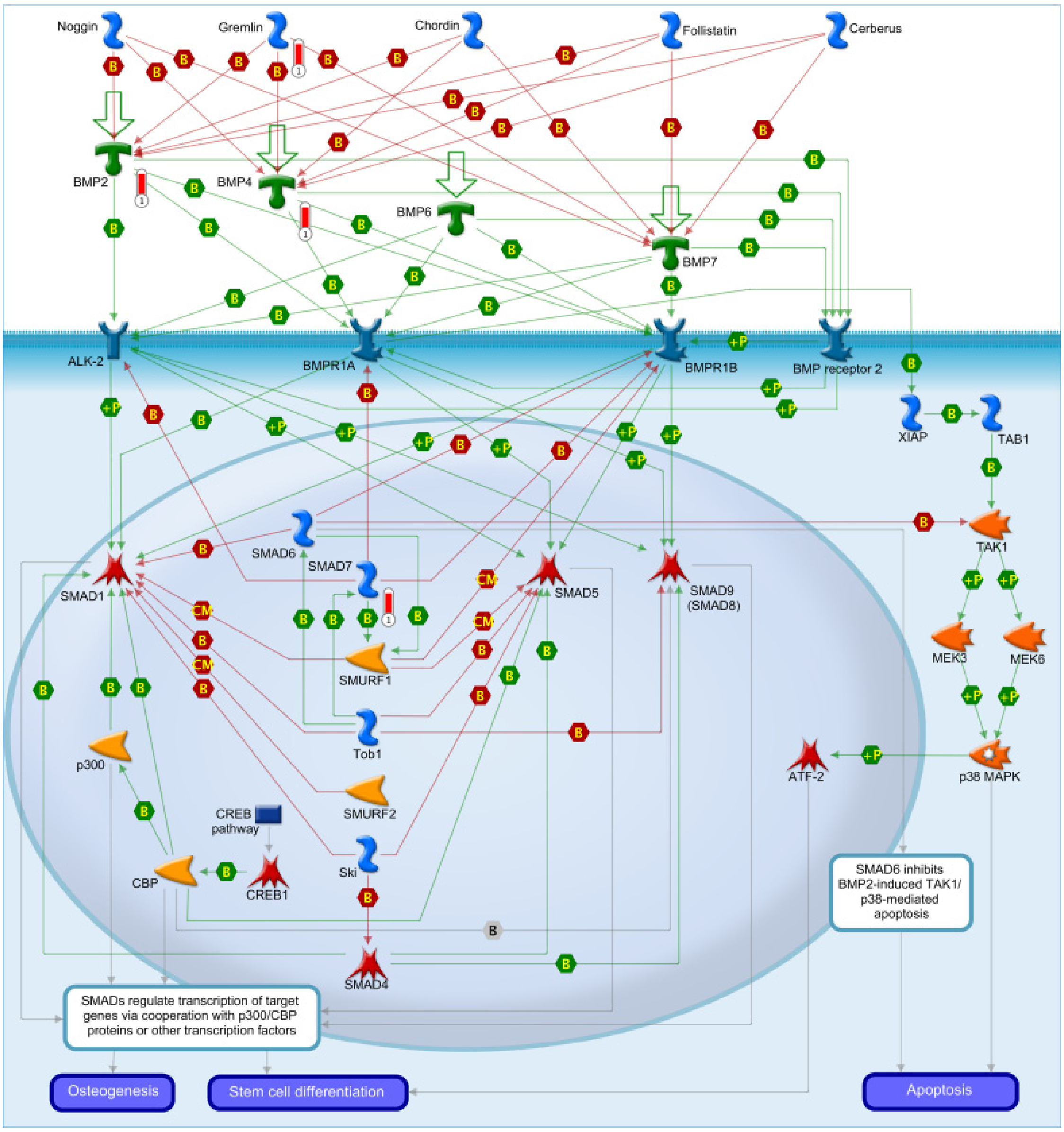


Figure 9.3 BMP signalling pathway



9.6 The screening of CDH1

Table 9.1 The forward and reverse primers used for the amplification of the CDH1 exon sequences

The label ‘SEQ’ under Light Scanner conditions indicates that the screening was performed by sequencing instead of the light scanner, this was performed for all fragments containing more than one SNP and any that were produced ambiguous results.

Primer ID	Primer Sequence (5’to 3’)	Stock dilution (mM)	Product Size (bp)	Exonic SNPs	PCR conditions		
					Temp (°C)	MgCl$_2$ (µg/µl)	Light-Scanner
CDH1_EX1FW	AGCACCTGTGAGCTTGC	200	253	0	55	2.5, Q	2.5Mg
CDH1_EX1REV	AGAAGGGAAGCGGTGAC						
CDH1_EX2FW	GTTTCGGTGAGCAGGAG	200	248	0	60	1.5	SEQ
CDH1_EX2REV	GGAGTGCAATTTCTCGG						
CDH1_EX3FW	GCTCTTTGGAGAAGGAATG	200	356	2	55	2.5	SEQ
CDH1_EX3REV	AAACCTGGATTAGACAGCG						
CDH1_EX4FW	GACCTGAAGTATCCGTCTTG	200	259	1	55	1.5	2.5Mg
CDH1_EX4REV	TCCTTGGTACTTCTCTGCC						
CDH1_EX5FW	AGTACCAAGGAGAGAAAGGG	200	289	0	55	1.5	2.5Mg +Q
CDH1_EX5REV	AAAATCCTGGGTGGATG						
CDH1_EX6FW	CTCAGAGCCTAGGAAGGTG	200	319	0	55	2.5	2.5Mg +Q
CDH1_EX6REV	CCAAGAAGTTCTGTCCGTAG						
CDH1_EX7FW	TTGACCCAGTCCCAAAG	200	314	1	55	2.5	2.5Mg
CDH1_EX7REV	TAGCAGGATTTTGCTTTGTC						
CDH1_EX8FW	CCAAAGGTGGCTAGTGTTC	200	265	0	55	2.5	2.5Mg
CDH1_EX8REV	CCATGAGCAGTGGTGAC						
CDH1_EX9FW	GAGGAATCCTTTAGCCCC	200	517	2	55	2.5	SEQ
CDH1_EX9REV	AGAAGATACCAGGGGACAAG						
CDH1_EX10FW2	CAAAAGCAACAGTTAAGGAT	200	489	2	55	1.5	SEQ
CDH1_EX10REV2	GAAAGGAGCACAGATAAAGG						
CDH1_EX11FW	TTCAGCTACATGTTGTTTGC	200	282	1	55	2.5	2.5Mg +Q
CDH1_EX11REV	TCCAAAAGAAGGGAGGG						
CDH1_EX12FW	TAGACTTGGTCTGGTGGAAG	200	356	5	60 L	3	SEQ
CDH1_EX12REV	GAAGGGAAGCATGGCAG						
CDH1_EX13FW	GGGTGTCTTTAGTTCACTAGC	200	392	2	55	2.5	SEQ
CDH1_EX13REV	TCCAGGAAATAAACCTCCTC						
CDH1_EX14FW	GAGGGGTGCTCTGTGATAG	200	271	1	55	2.5	2.5 +Q
CDH1_EX14REV	TGCTTCTTCCGAATAAAGAG						
CDH1_EX15FW	AGTGAAGGCATCATCCAAC	200	370	0	55	2.5	2.5Mg +Q
CDH1_EX15REV	CATAGTAAAGGAAAGAATCTAAAGAC						
CDH1_EX16FW	TATTGCTAGACTTCTTGCCC	200	375	2	55 L	1.5	SEQ
CDH1_EX16REV	AAACTCATCTCAAGGGAAGG						

9.7 The replication of the associated X chromosome SNPs

Table 9.2 Kaspar Primer Sequences

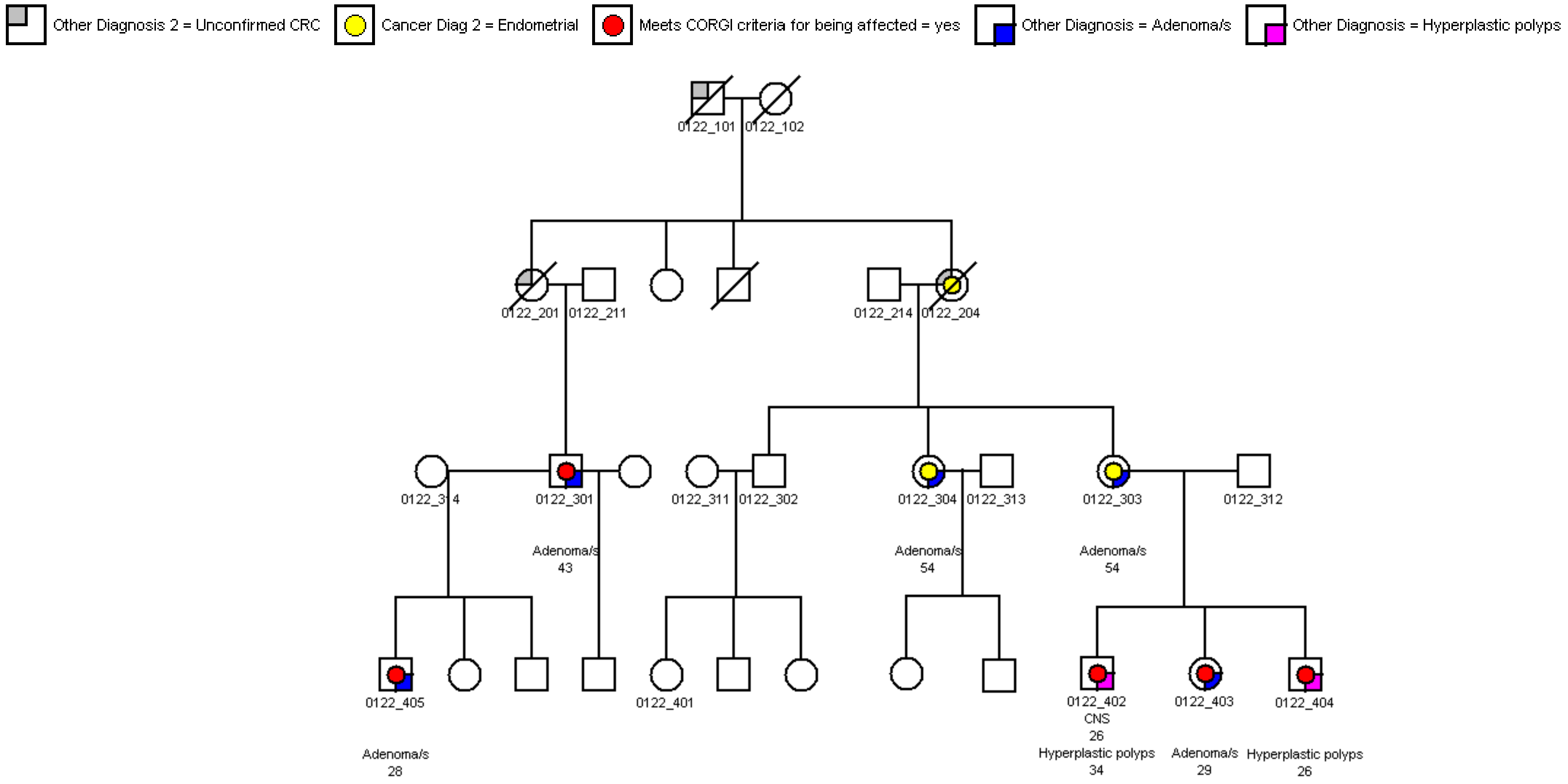
The primer sequences for the associated SNPs identified on the X chromosome, which were genotyped in CORGI2bcd using Kaspar. This technique involves three primers, one that is specific for each allele and a common reverse primer

SNP	Primer	Primer Sequence	Direction
rs12860832	A allele	GAAGGTGACCAAGTTCATGCTCATAAAATTTGCAGTATGCTGAGTTGGT	Reverse
	G allele	GAAGGTCGGAGTCAACGGATTATAAAATTTGCAGTATGCTGAGTTGGC	Reverse
	Common	CAGGACTCTGAAATCCTTCCTTCCAA	Reverse
rs5934683	C allele	GAAGGTGACCAAGTTCATGCTTCTGAAAATTCCACCTGAGC	Forward
	T allele	GAAGGTCGGAGTCAACGGATTCTGCTTCTGAAAATTCCACCTGAGT	Forward
	Common	GTGTATGGACTCCTAGTAGATGGCTT	Forward

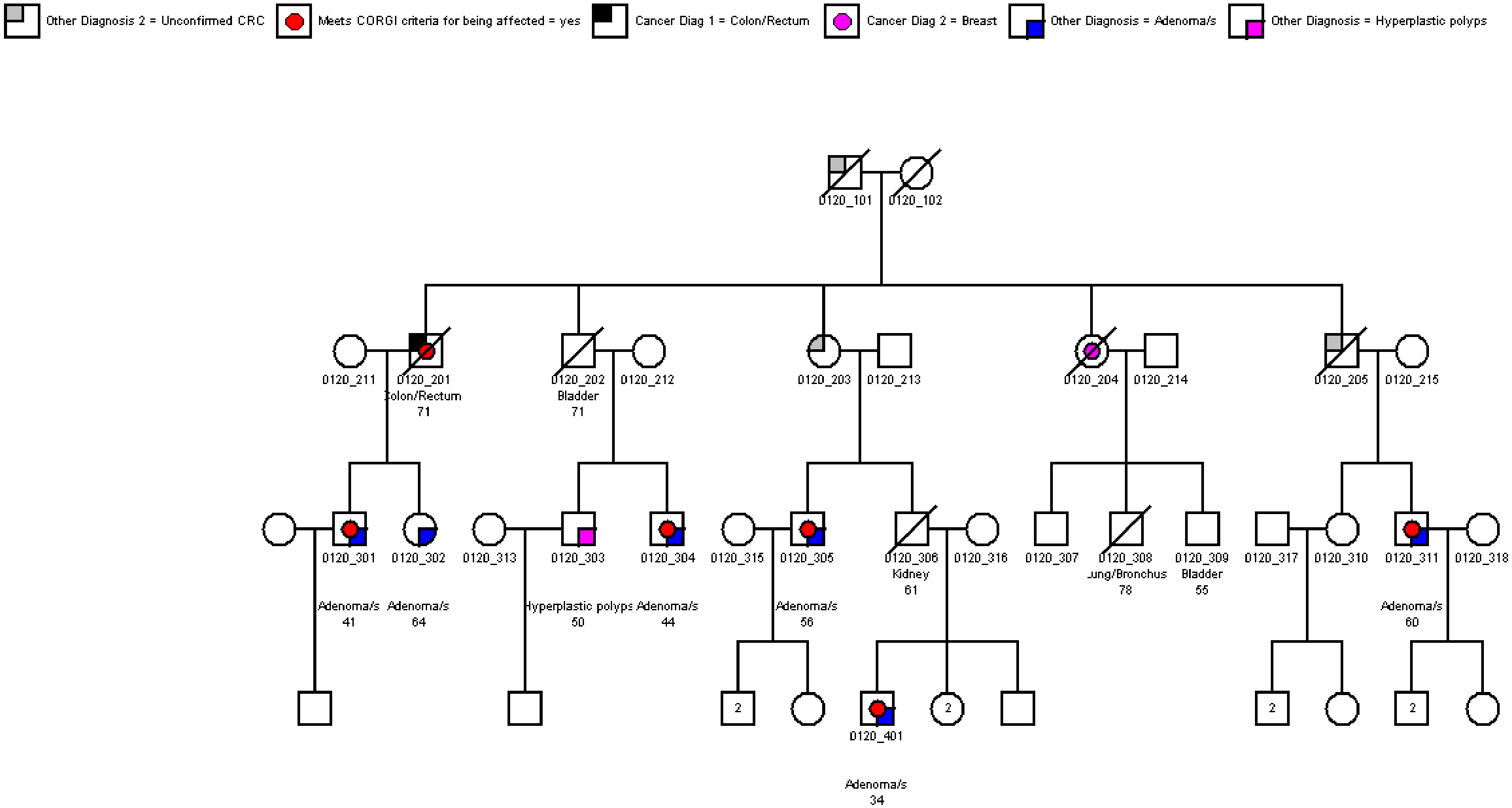
9.8 The detection of moderate penetrance, rare susceptibility alleles

Figure 9.4 The pedigree for the CORGI families included in the study

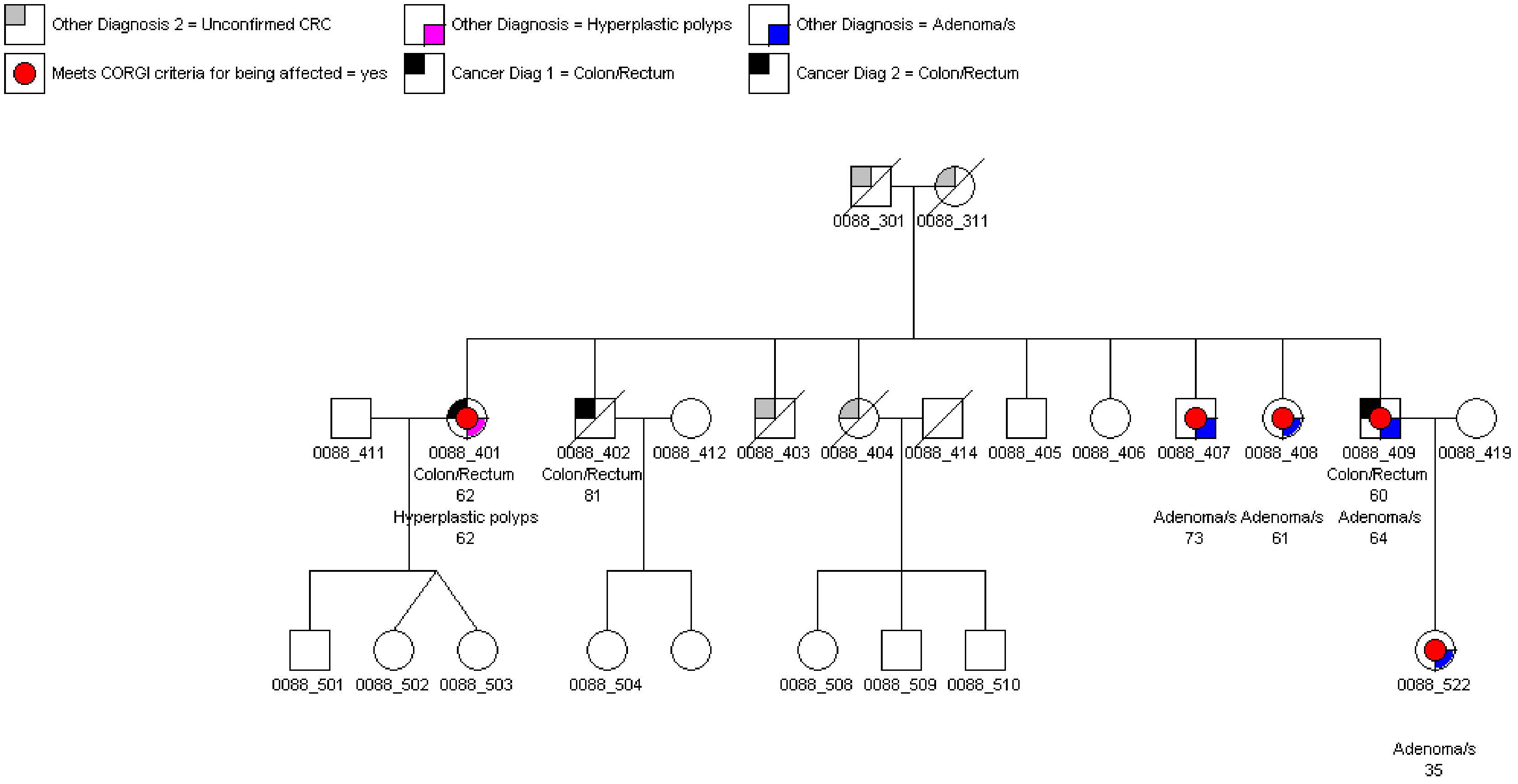
Family 336



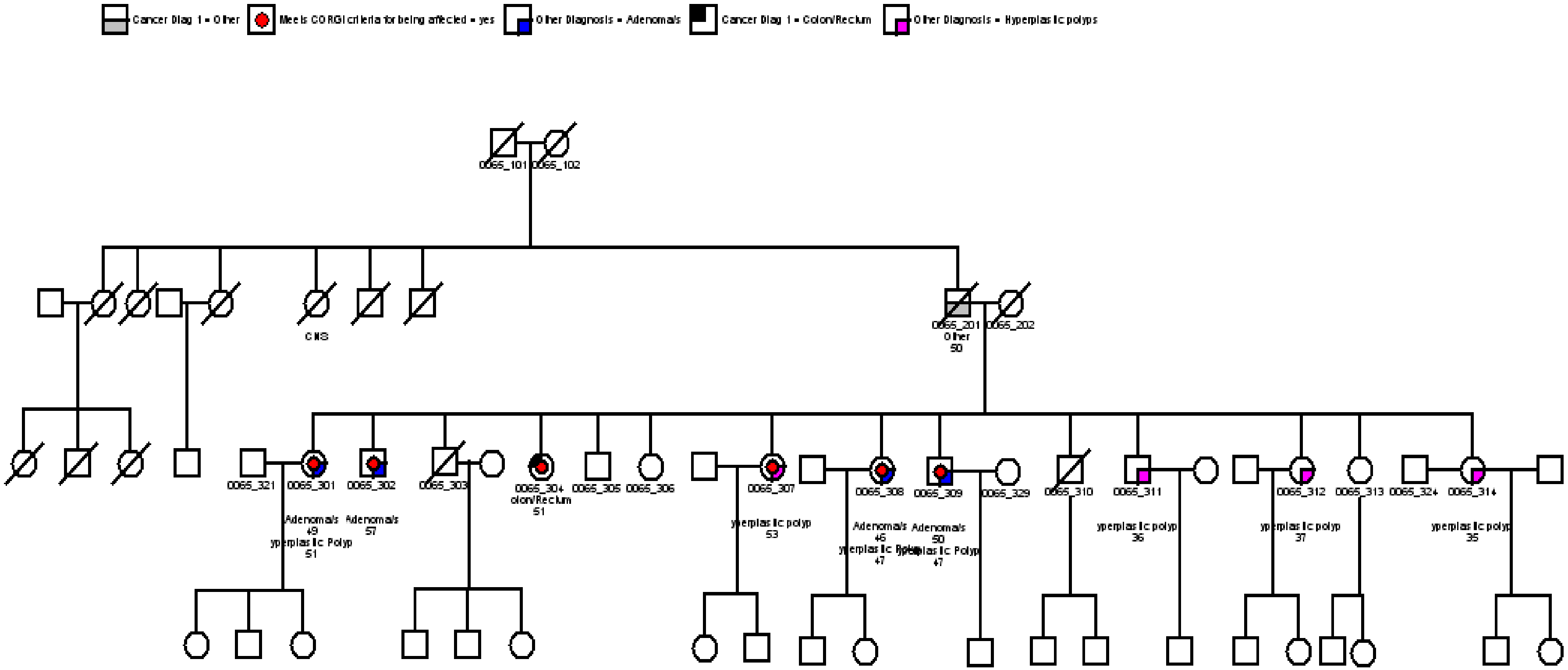
Family 323


Other Diagnosis 2 = Unconfirmed CRC
Meets CORGI criteria for being affected = yes
Cancer Diag 1 = Colon/Rectum
Cancer Diag 2 = Breast
Other Diagnosis = Adenoma/s
Other Diagnosis = Hyperplastic polyps
0120_101
0120_102
0120_211
0120_201
Colon/Rectum
71
0120_202
Bladder
71
0120_212
0120_203
0120_213
0120_204
0120_214
0120_205
0120_215
0120_301
Adenoma/s
41
0120_302
Adenoma/s
64
0120_313
0120_303
Hyperplastic polyps
50
0120_304
Adenoma/s
44
0120_315
0120_305
Adenoma/s
56
0120_306
Kidney
61
0120_316
0120_307
0120_308
Lung/Bronchus
78
0120_309
Bladder
55
0120_317
0120_310
0120_311
Adenoma/s
60
0120_318
2
2
0120_401
Adenoma/s
34
2
2

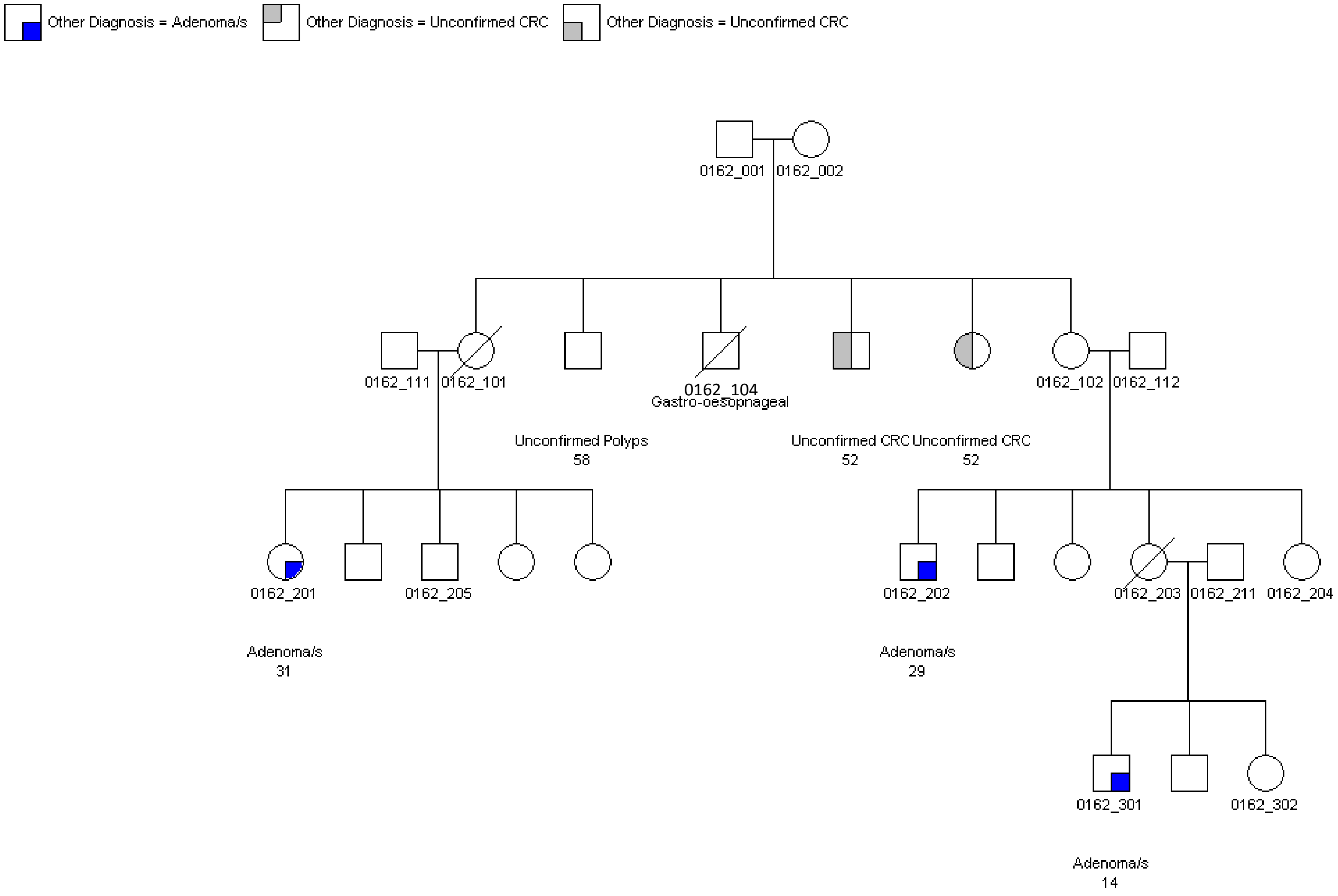
Family 282


Other Diagnosis 2 = Unconfirmed CRC
Other Diagnosis = Hyperplastic polyps
Other Diagnosis = Adenoma/s
Meets CORGI criteria for being affected = yes
Cancer Diag 1 = Colon/Rectum
Cancer Diag 2 = Colon/Rectum
0088_301
0088_311
0088_411
0088_401
Colon/Rectum
62
Hyperplastic polyps
62
0088_402
Colon/Rectum
81
0088_412
0088_403
0088_404
0088_414
0088_405
0088_406
0088_407
Adenoma/s
73
0088_408
Adenoma/s
61
0088_409
Colon/Rectum
60
Adenoma/s
64
0088_419
0088_501
0088_502
0088_503
0088_504
0088_508
0088_509
0088_510
0088_522
Adenoma/s
35

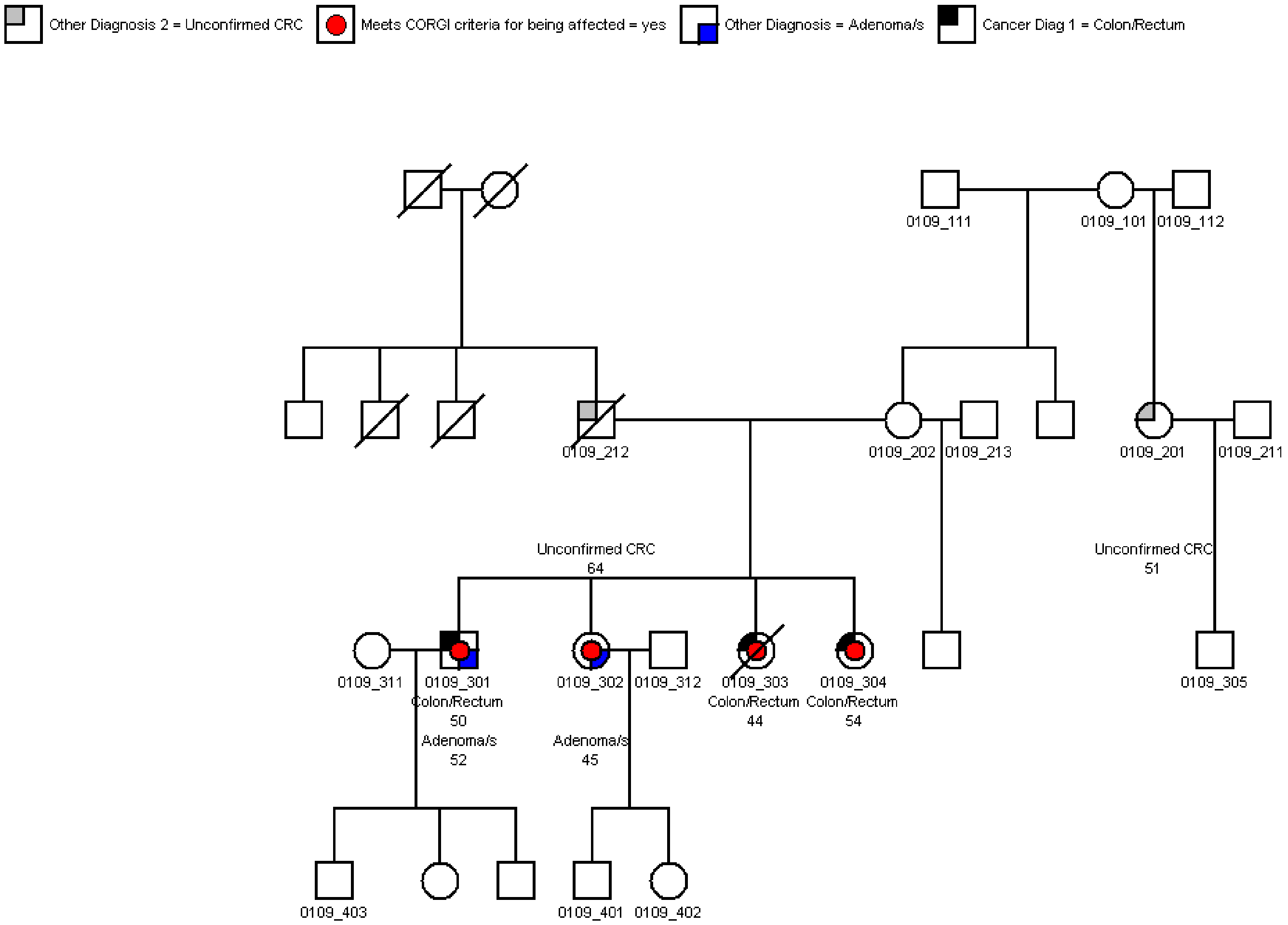
Family 294


Cancer Diag 1 = Other
Meets CORGI criteria for being affected = yes
Other Diagnosis = Adenomals
Cancer Diag 1 = Colon/Rectum
Other Diagnosis = Hyperplastic polyps
0065_101
0065_102
CNS
0065_201
Other
50
0065_202
0065_321
0065_301
Adenomals
49
yperplastic Polyp
51
0065_302
Adenomals
57
0065_303
0065_304
olon/Rectum
51
0065_305
0065_306
0065_307
yperplastic polyp
53
0065_308
Adenomals
46
yperplastic Polyp
47
0065_309
Adenomals
50
yperplastic Polyp
47
0065_329
0065_310
0065_311
yperplastic polyp
36
0065_312
yperplastic polyp
37
0065_313
0065_324
0065_314
yperplastic polyp
35

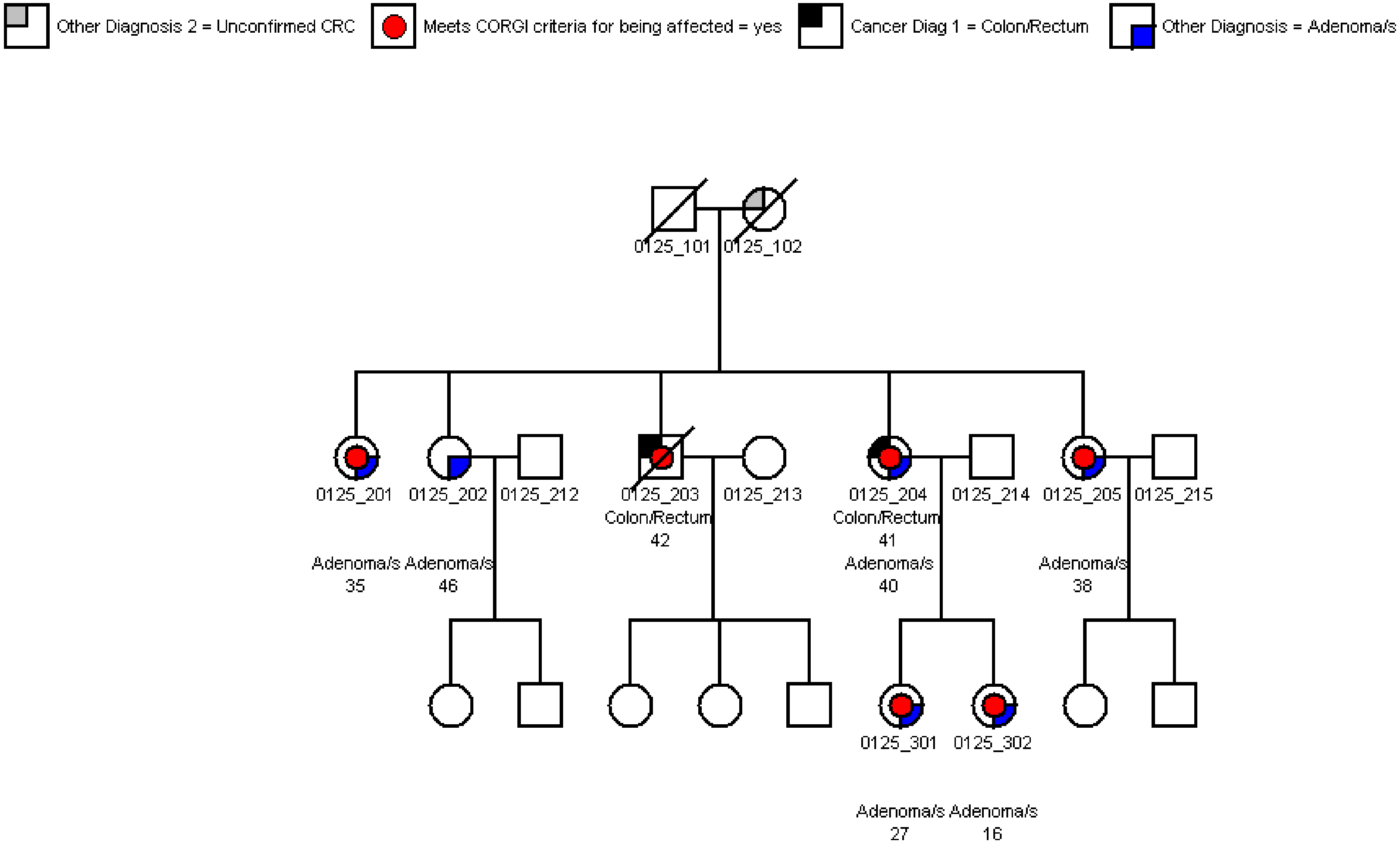
Family 450 – an additional family that was included in the LOH, but not the linkage analysis.



Family 329


Other Diagnosis 2 = Unconfirmed CRC
Meets CORGI criteria for being affected = yes
Other Diagnosis = Adenoma/s
Cancer Diag 1 = Colon/Rectum
0109_111
0109_101
0109_112
0109_212
0109_202
0109_213
0109_201
0109_211
Unconfirmed CRC
64
Unconfirmed CRC
51
0109_311
0109_301
Colon/Rectum
50
Adenoma/s
52
0109_302
Adenoma/s
45
0109_312
0109_303
Colon/Rectum
44
0109_304
Colon/Rectum
54
0109_305
0109_403
0109_401
0109_402

Family 377


Other Diagnosis 2 = Unconfirmed CRC
Meets CORGI criteria for being affected = yes
Cancer Diag 1 = Colon/Rectum
Other Diagnosis = Adenoma/s
0125_101
0125_102
0125_201
Adenoma/s
35
0125_202
Adenoma/s
46
0125_212
0125_203
Colon/Rectum
42
0125_213
0125_204
Colon/Rectum
41
Adenoma/s
40
0125_214
0125_205
Adenoma/s
38
0125_215
0125_301
Adenoma/s
27
0125_302
Adenoma/s
16

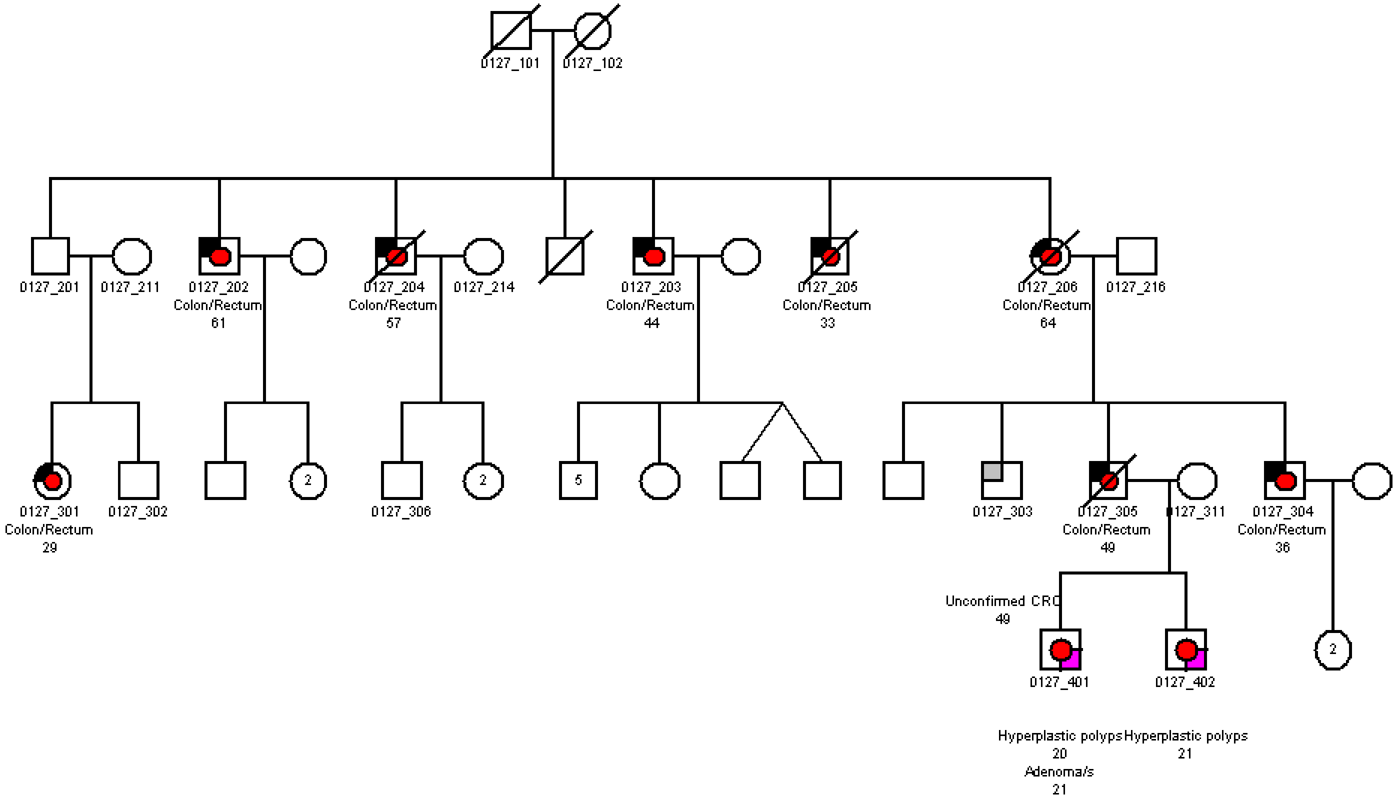
Family 346

● Meets CORGI criteria for being affected = yes

◩ Cancer Diag 1 = Colon/Rectum

◩ Other Diagnosis 2 = Unconfirmed CRC

◪ Other Diagnosis = Hyperplastic polyps



Family 326

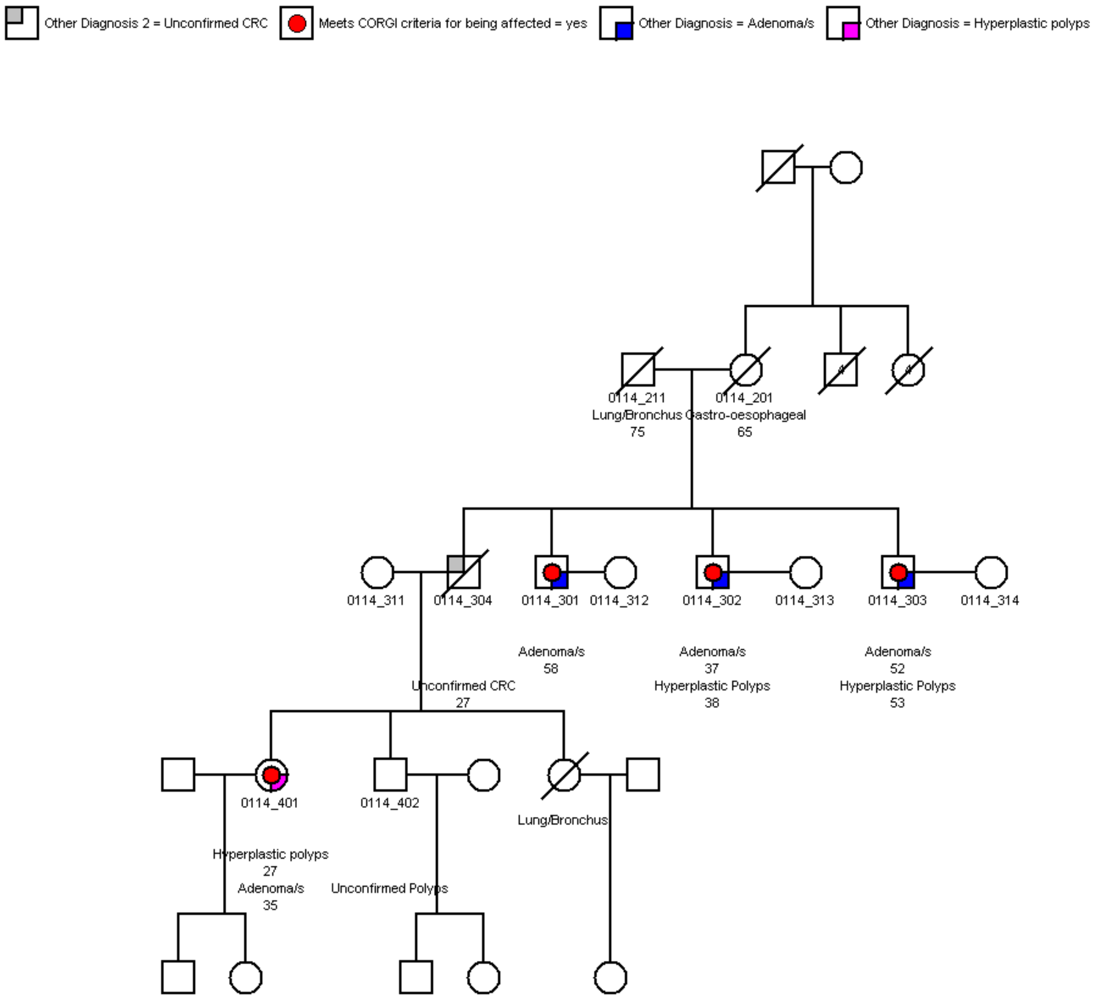
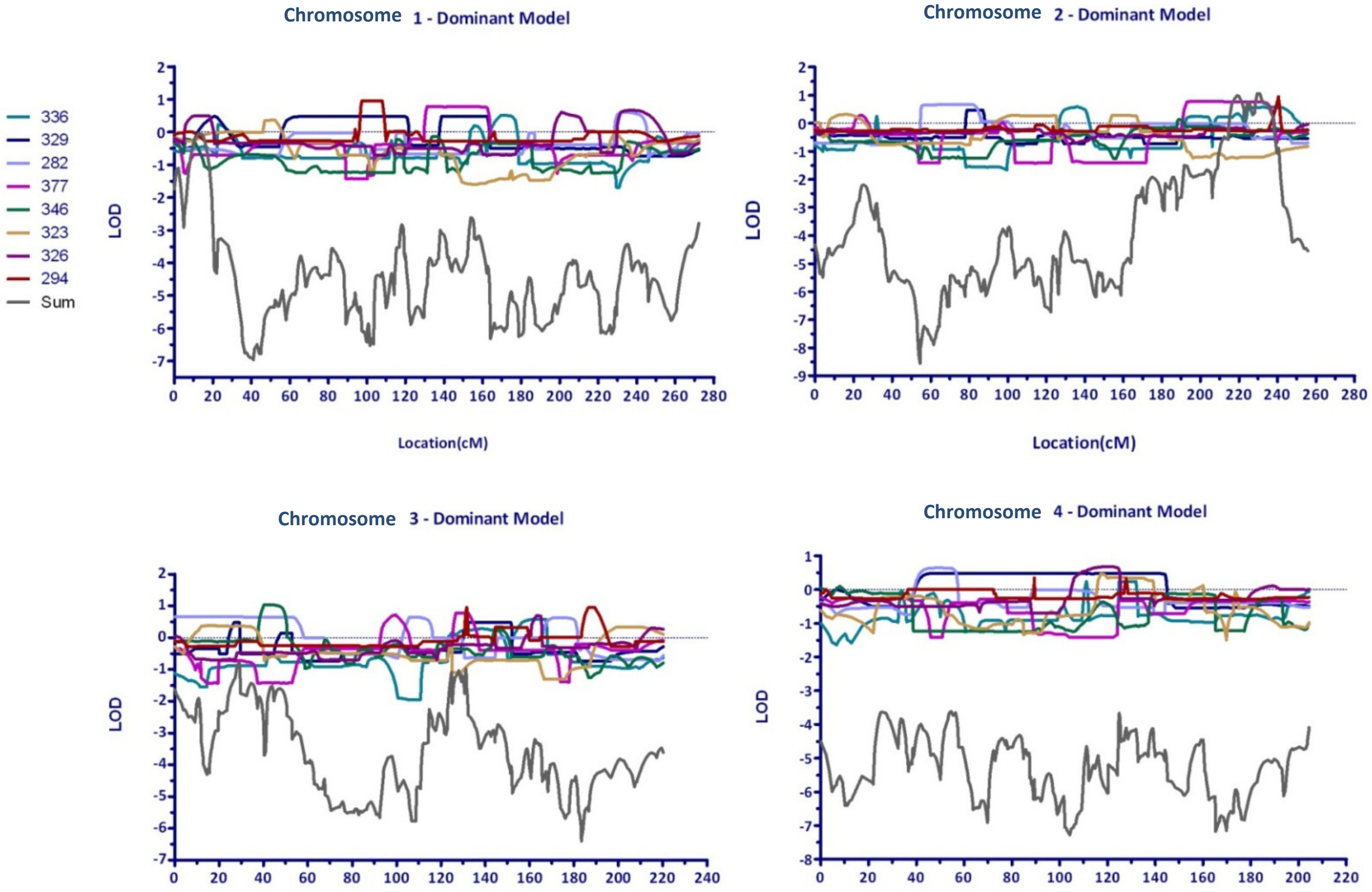

Other Diagnosis 2 = Unconfirmed CRC
Meets CORGI criteria for being affected = yes
Other Diagnosis = Adenoma/s
Other Diagnosis = Hyperplastic polyps
0114_211
Lung/Bronchus
75
0114_201
Gastro-oesophageal
65
4
4
0114_311
0114_304
Unconfirmed CRC
27
0114_301
Adenoma/s
58
0114_312
0114_302
Adenoma/s
37
Hyperplastic Polyps
38
0114_313
0114_303
Adenoma/s
52
Hyperplastic Polyps
53
0114_314
0114_401
Hyperplastic polyps
27
Adenoma/s
35
0114_402
Unconfirmed Polyps
Lung/Bronchus

Figure 9.5 Dominant model genome-wide linkage analysis results showing all chromosomes not given in the chapter 7

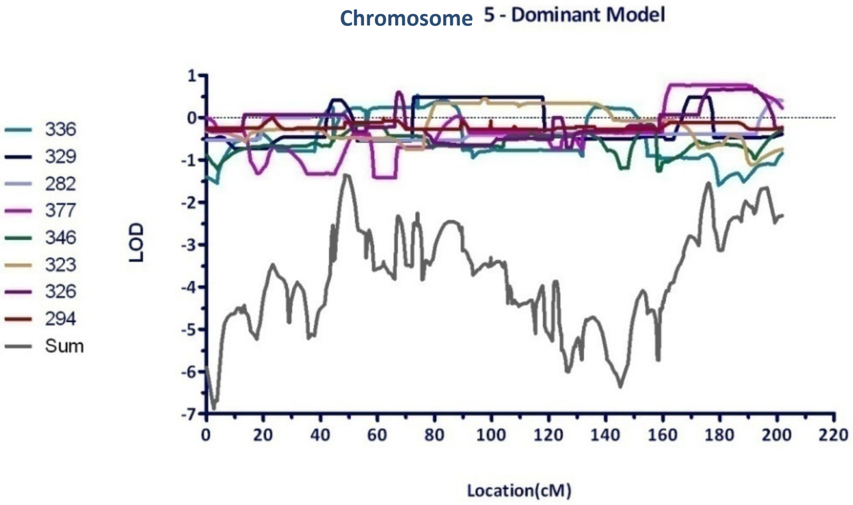


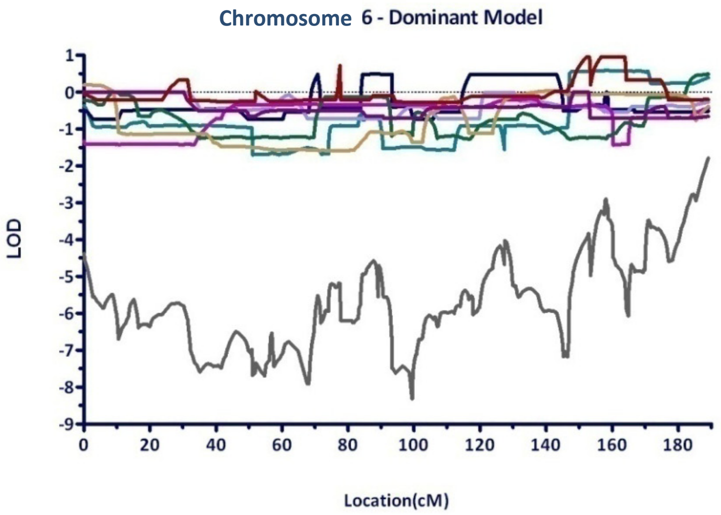
Chromosome 1 - Dominant Model

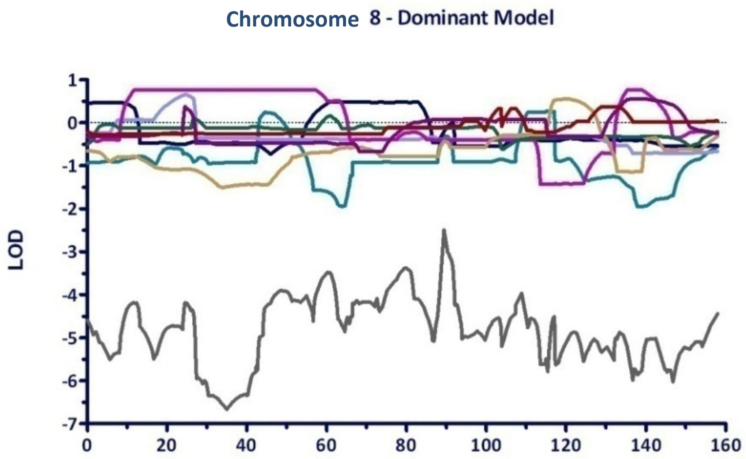
Chromosome 2 - Dominant Model

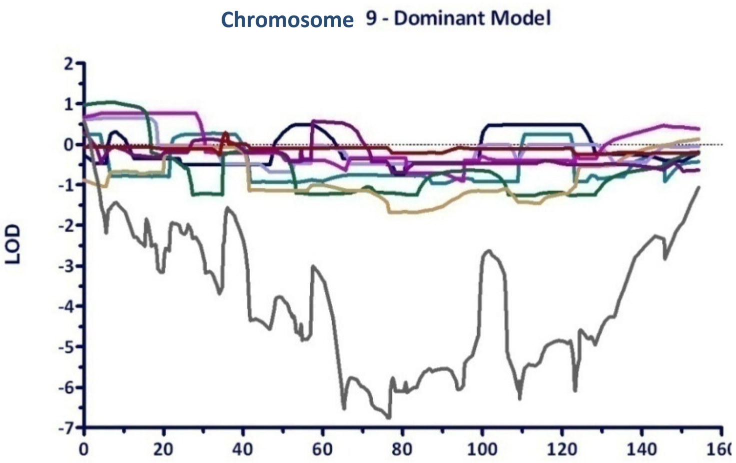
Chromosome 3 - Dominant Model

Chromosome 4 - Dominant Model

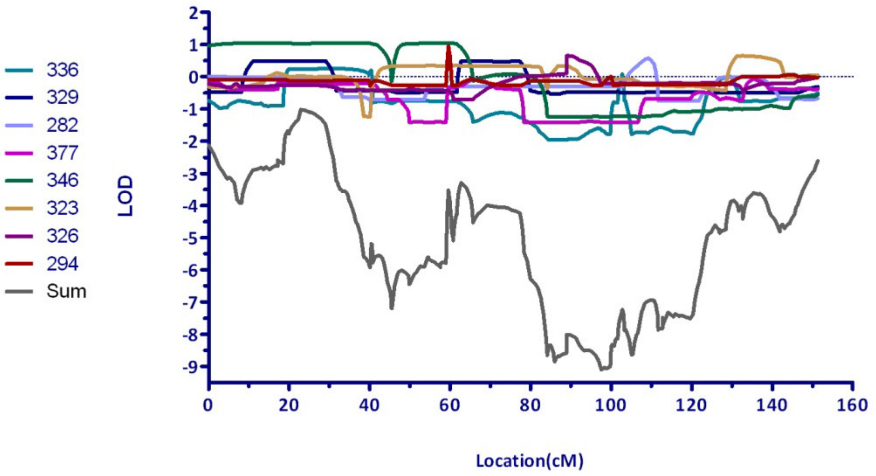

Chromosome 5 - Dominant Model


Chromosome 6 - Dominant Model

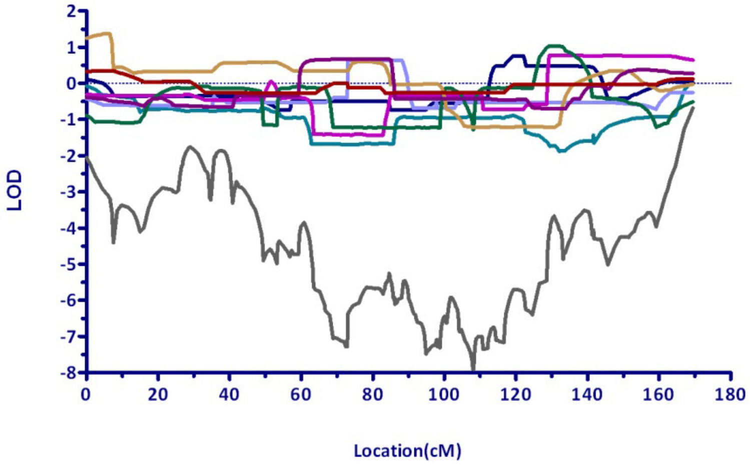

Chromosome 8 - Dominant Model


Chromosome 9 - Dominant Model

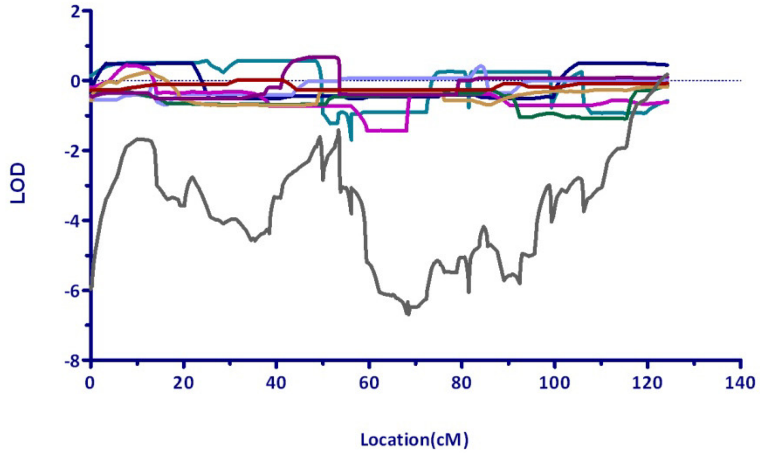
Chromosome 11 - Dominant Model



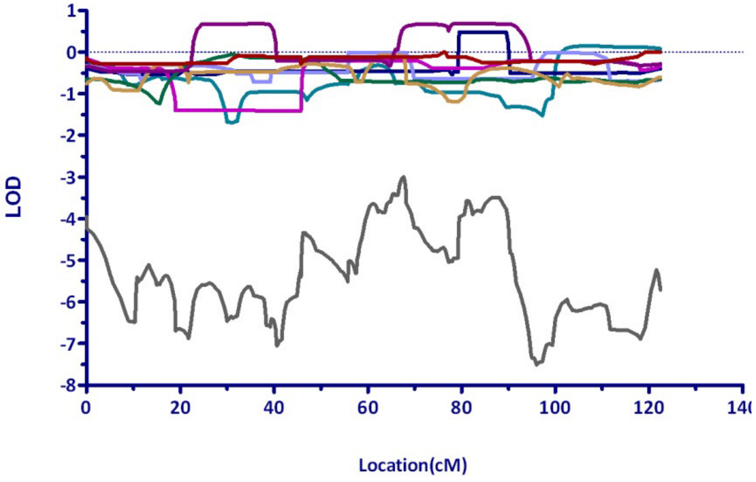
Chromosome 12 - Dominant Model

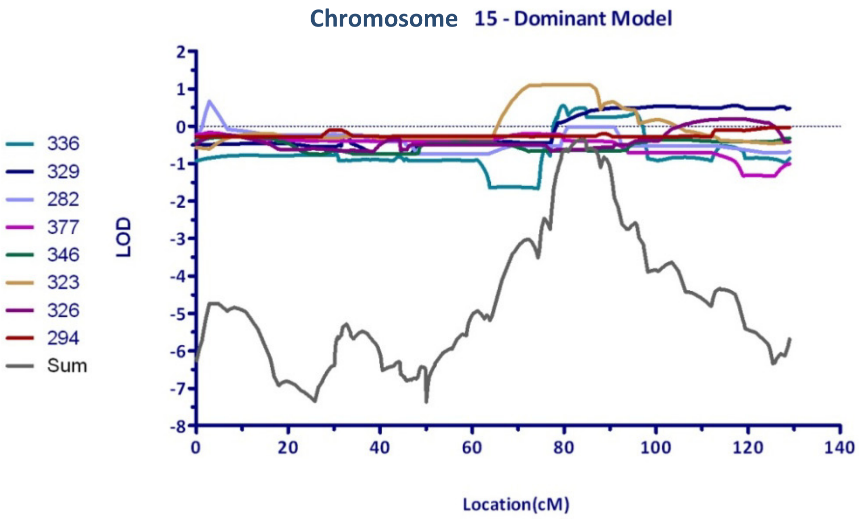


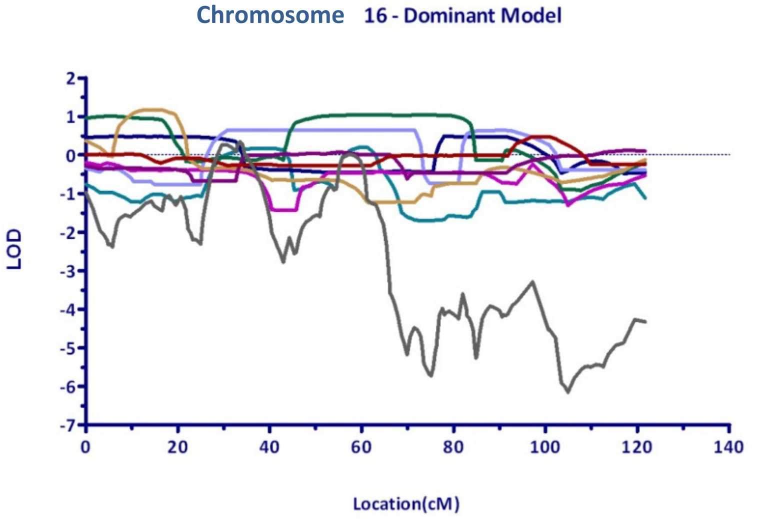
Chromosome 13 - Dominant Model

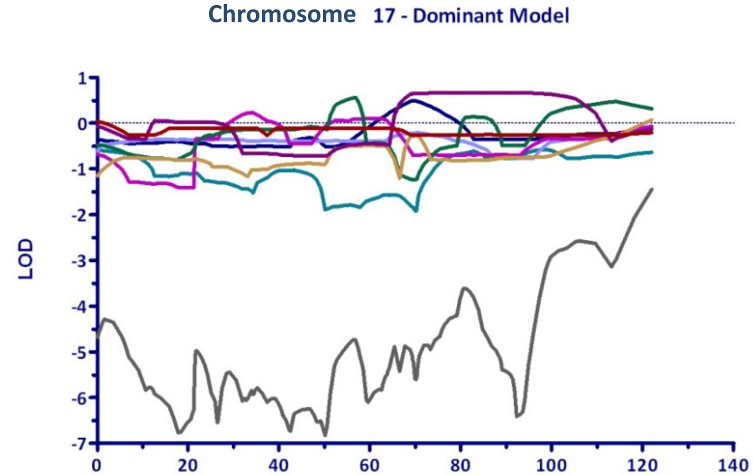


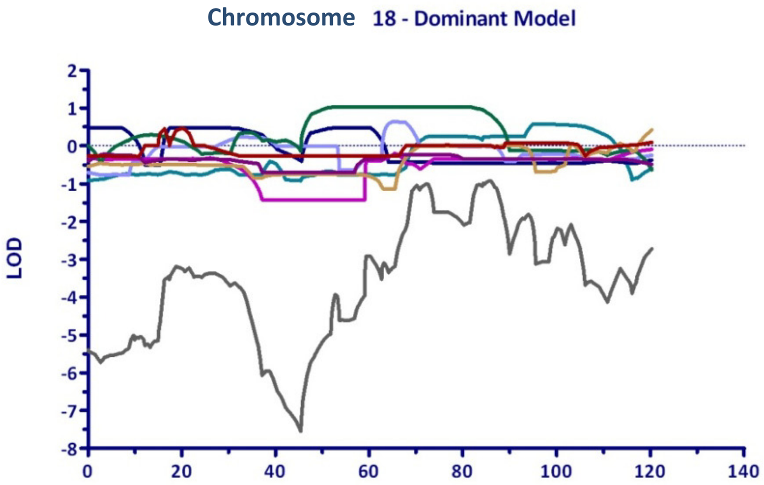
Chromosome 14 - Dominant Model

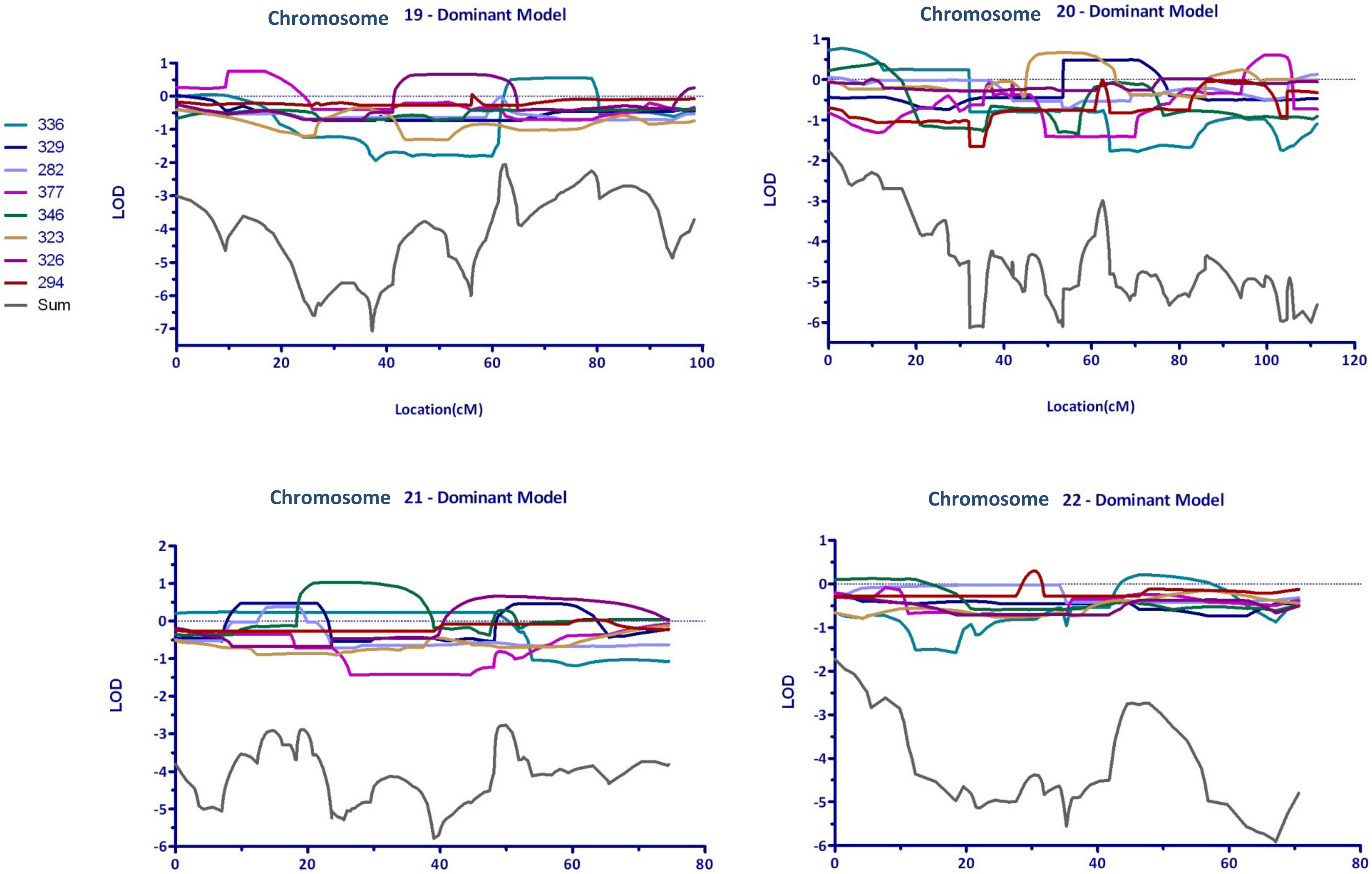



Chromosome 15 - Dominant Model


Chromosome 16 - Dominant Model


Chromosome 17 - Dominant Model


Chromosome 18 - Dominant Model



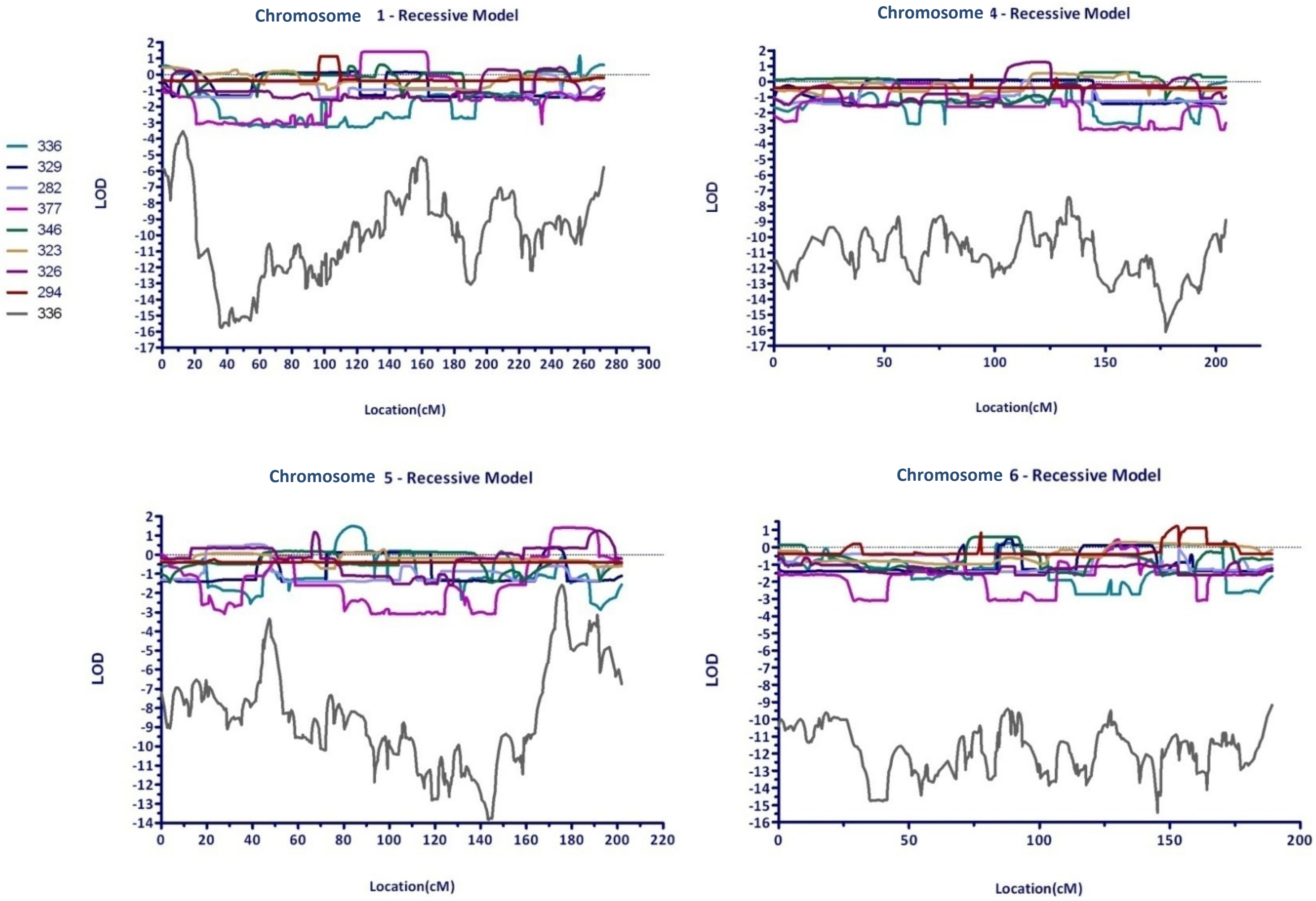
Chromosome 19 - Dominant Model

Chromosome 20 - Dominant Model

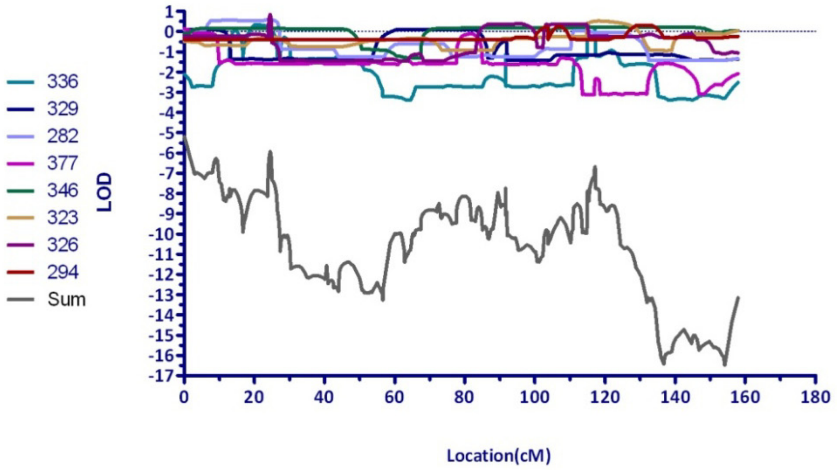
Chromosome 21 - Dominant Model

Chromosome 22 - Dominant Model

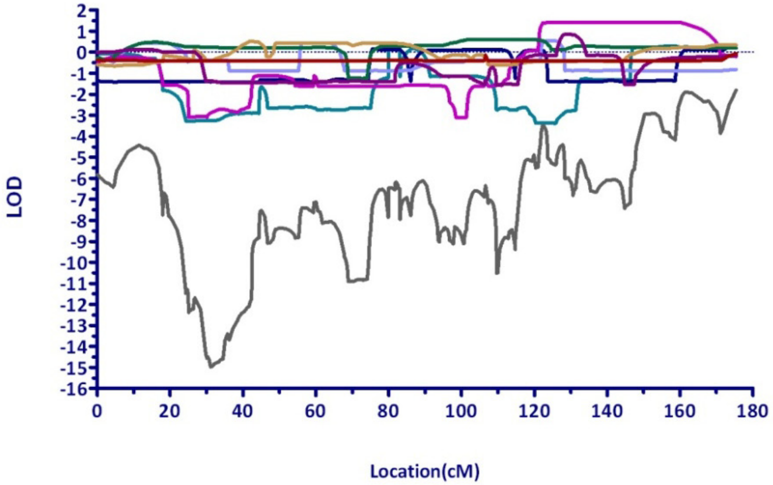
Figure 9.6 Recessive model genome -wide linkage analysis results showing all chromosomes not given in the chapter 7



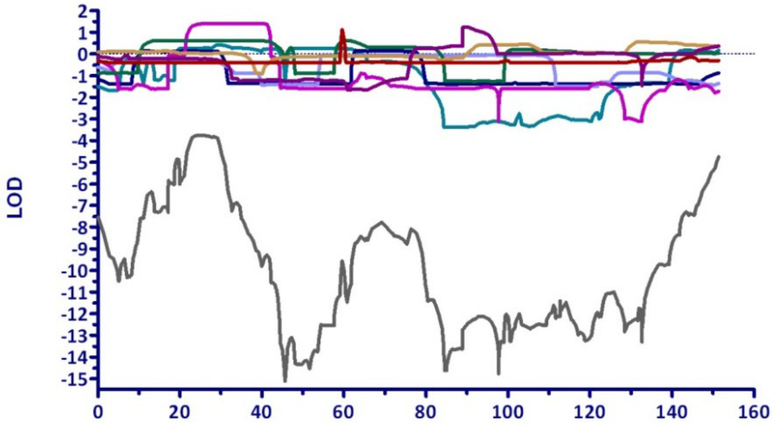
Chromosome 8 - Recessive Model



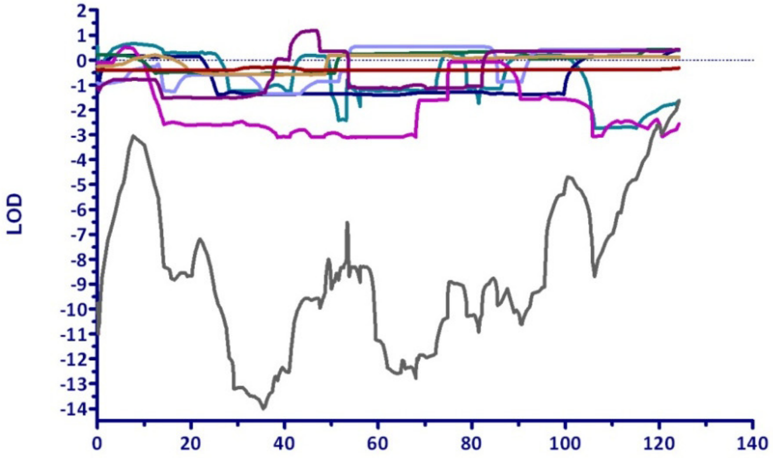
Chromosome 10 - Recessive Model

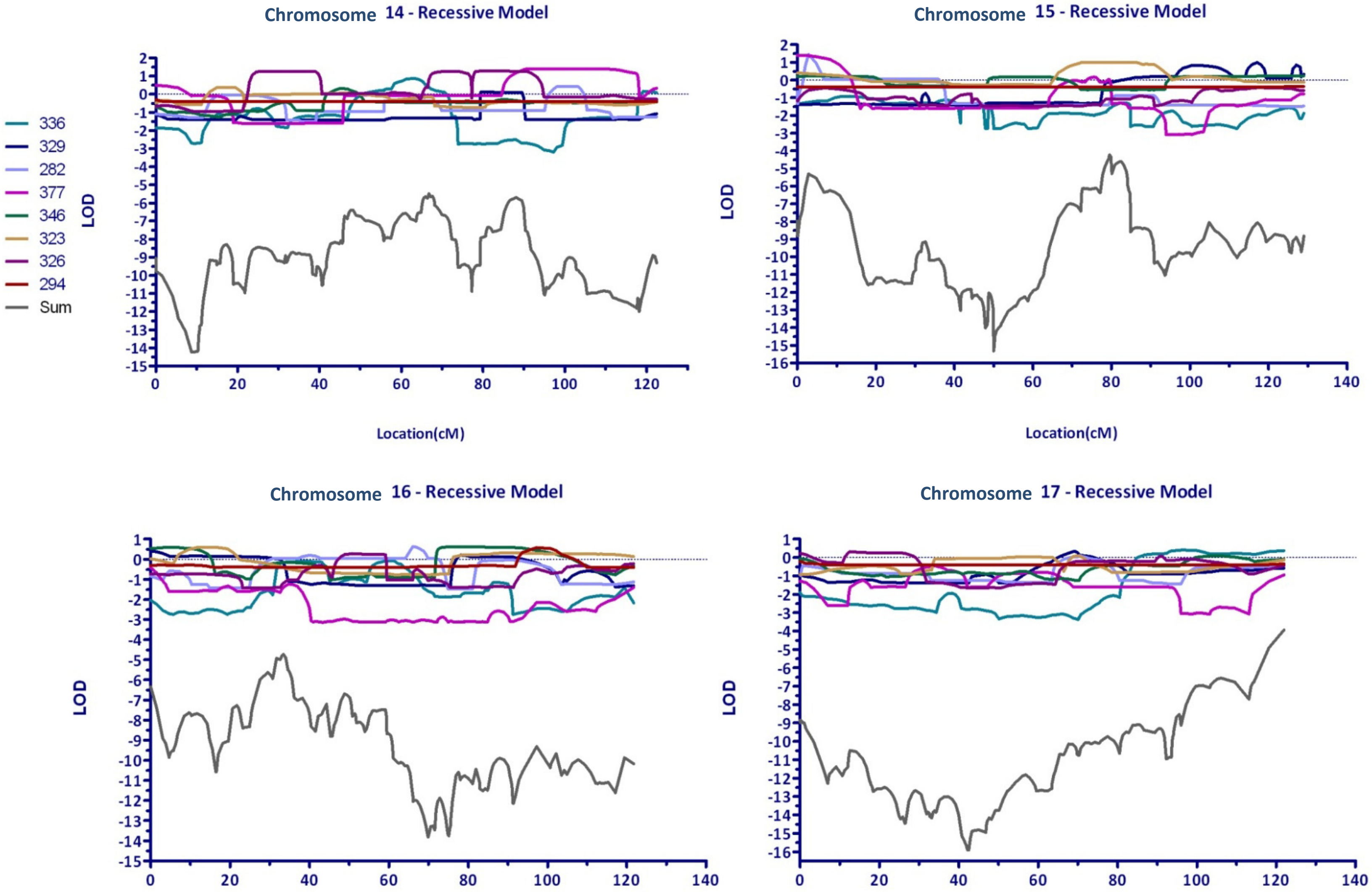


Chromosome 11 - Recessive Model

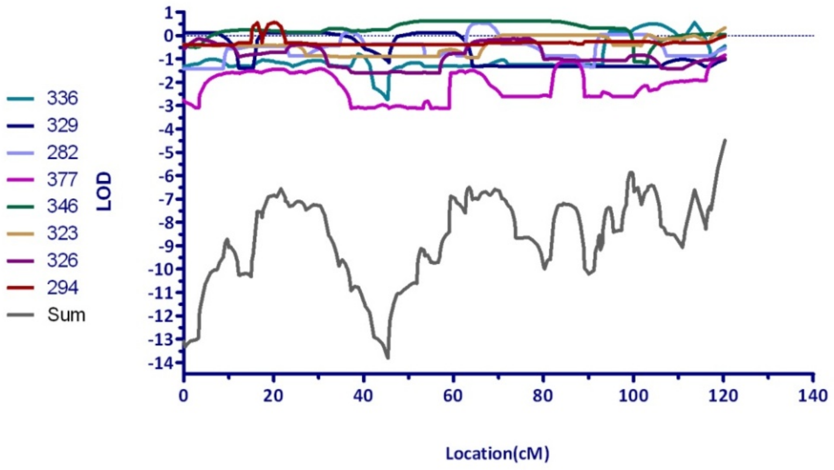


Chromosome 13 - Recessive Model

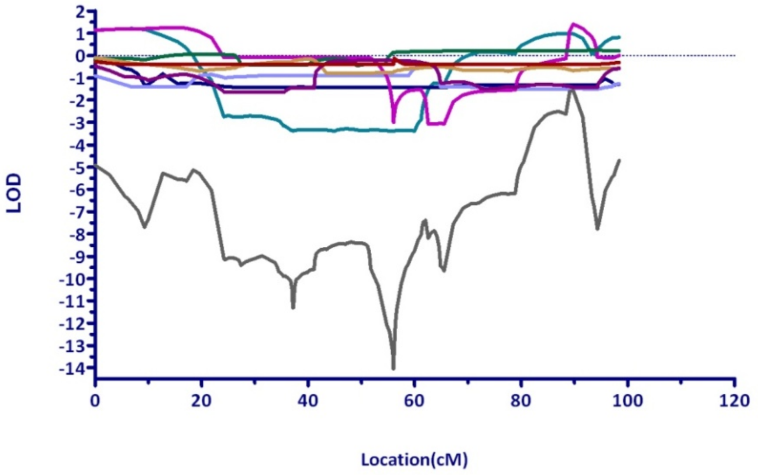



Chromosome 14 - Recessive Model
Chromosome 15 - Recessive Model
Chromosome 16 - Recessive Model
Chromosome 17 - Recessive Model
336
329
282
377
346
323
326
294
Sum
LOD
Location(cM)
2
1
0
-1
-2
-3
-4
-5
-6
-7
-8
-9
-10
-11
-12
-13
-14
-15
-16
0
20
40
60
80
100
120
140

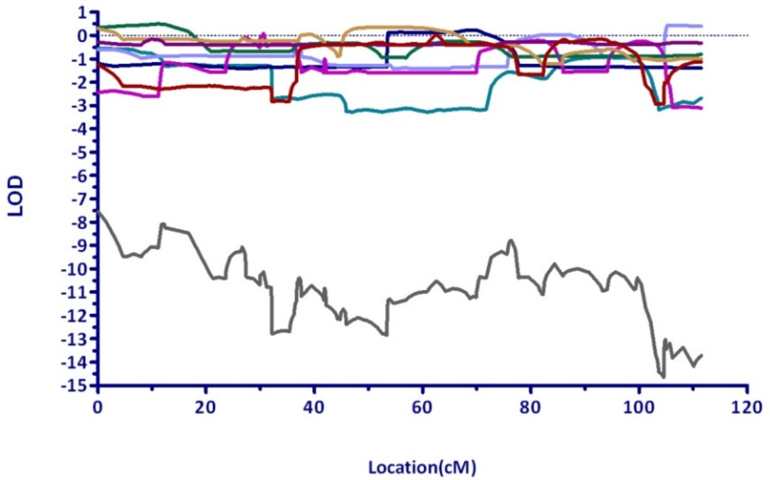
Chromosome 18 - Recessive Model



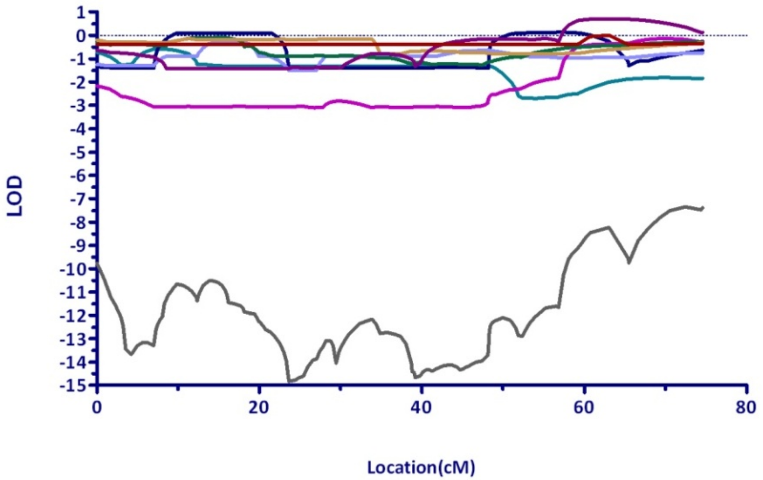
Chromosome 19 - Recessive Model



Chromosome 20 - Recessive Model



Chromosome 21 - Recessive Model



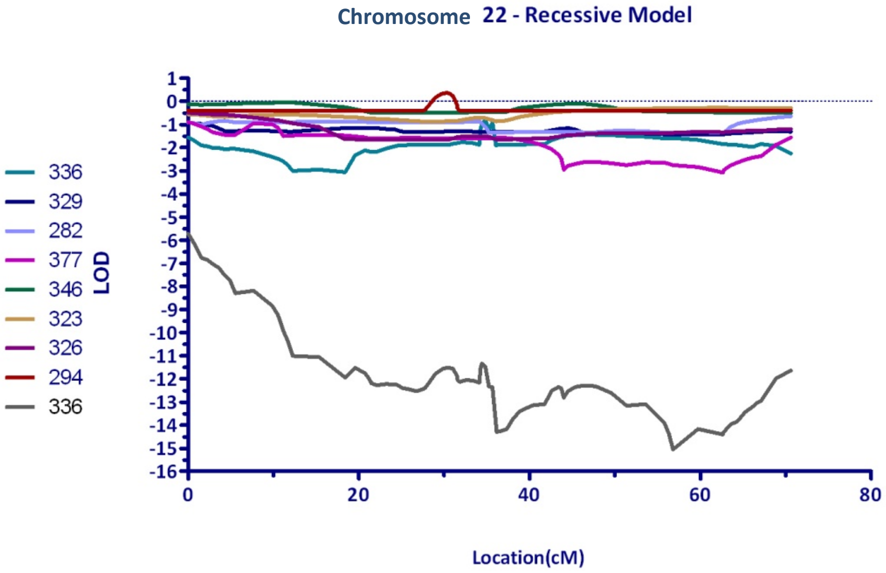

Chromosome 22 - Recessive Model
336
329
282
377
346
323
326
294
336
LOD
1
0
-1
-2
-3
-4
-5
-6
-7
-8
-9
-10
-11
-12
-13
-14
-15
-16
0
20
40
60
80
Location(cM)

Table 9.3 The common regions of LOH detected within families

This table provides the full details of the ROHs detected that contribute to the common regions identified within certain families and discussed in sections 1.3.3.1 and 1.3.3.2. The group column (GRP) indicates the grouping of segments that make up each pool based on segments with 95% allelic identity of the genotypes. A "*" indicates the reference sample.

POOL	IID	Tumour ID	Chr.	SNP1	SNP2	Start (bp)	End (bp)	Size (kb)	SNPs	NSIM	GRP
S13	0120_304	0013077_1A	15	rs1869907	rs16952667	39,044,058	42,884,027	3839.97	12	5	1
S13	0120_304	0013077_4A	15	rs1869907	rs16952667	39,044,058	42,884,027	3839.97	12	5	1
S13	0120_301	04_12868_2A_2	15	rs10520142	rs877007	37,729,189	47,116,195	9387.01	32	5	1
S13	0120_301	04_12868_2B	15	rs10520142	rs3198	37,729,189	45,473,555	7744.37	25	5	1
S13	0120_301	04_12868_2A_1	15	rs10520142	rs3198	37,729,189	45,473,555	7744.37	25	5	1
S13	0120_301	04_12868_2C	15	rs276855	rs1048975	37,318,605	47,204,838	9886.23	34	5	1*
S13	0120_311	0206460_1A	15	rs1433887	rs16952667	37,016,395	42,884,027	5867.63	20	0	2*
S13	0120_305	0011426_1A	15	rs1565863	rs1648282	38,225,412	43,213,156	4987.74	14	0	3*
S13	CON	8	15	rs1869907	rs16952667	39,044,058	42,884,027	3839.97	12	NA	NA
S13	UNION	8	15	rs1433887	rs1048975	37,016,395	47,204,838	10188.4	39	NA	NA
S75	0125_203	95_2882_C_N	17	rs4890140	rs962272	36,301,688	44,333,282	8031.59	17	19	1
S75	0125_204	97_13626_1A	17	rs4890140	rs962272	36,301,688	44,333,282	8031.59	17	19	1
S75	0125_302	SS02_037961A	17	rs4890140	rs962272	36,301,688	44,333,282	8031.59	17	11	1
S75	0125_302	SS02_037964A_C	17	rs4890140	rs962272	36,301,688	44,333,282	8031.59	17	11	1
S75	0125_302	SS02_037964A_D	17	rs4890140	rs962272	36,301,688	44,333,282	8031.59	17	11	1
S75	0125_302	044123_1A_D	17	rs1526601	rs1526189	36,088,662	47,211,844	11123.2	22	11	1
S75	0125_302	044123_1A_C	17	rs1526601	rs962272	36,088,662	44,333,282	8244.62	18	11	1
S75	0125_302	02/037964A/A	17	rs1526601	rs962272	36,088,662	44,333,282	8244.62	18	11	1
S75	0125_204	97/13626/2A	17	rs1526601	rs733920	36,088,662	43,994,702	7906.04	17	19	1
S75	0125_302	044123_1A_B	17	rs12600677	rs962272	29,958,091	44,333,282	14375.2	28	19	1*
S75	CON	20	17	rs1078830	rs2051821	41,301,901	43,498,514	2196.61	5	NA	NA

S75	UNION	20	17	rs12600677	rs2045418	29,958,091	49,458,230	19500.1	35	NA	NA
S114	0125_302	SS02_037961A	17	rs222850	rs1997095	7,078,674	8,915,247	1836.57	10	15	1
S114	0125_302	SS02_037964A_B	17	rs222850	rs1997095	7,078,674	8,915,247	1836.57	10	15	1
S114	0125_302	SS02_037964A_C	17	rs222850	rs1997095	7,078,674	8,915,247	1836.57	10	15	1
S114	0125_302	SS02_037964A_D	17	rs222850	rs1997095	7,078,674	8,915,247	1836.57	10	15	1*
S114	0125_302	044123_1A_D	17	rs12746	rs6503211	6,295,055	9,333,425	3038.37	16	7	1
S114	0125_302	044123_1A_C	17	rs405923	rs1997095	3,185,391	8,915,247	5729.86	24	7	1
S114	0125_302	02/037964A/A	17	rs1984749	rs6503211	2,196,654	9,333,425	7136.77	27	7	1
S114	0125_302	044123_1A_B	17	rs6502862	rs1266160	1,142,838	10,402,401	9259.56	37	7	1
S114	0125_201	00/03743/3A/C	17	rs1565816	rs995362	7,708,213	10,421,640	2713.43	14	12	2
S114	0125_201	S0110291/1A/C	17	rs858526	rs2240519	7,440,107	11,716,085	4275.98	19	12	2
S114	0125_201	00/03743/1A	17	rs858526	rs995362	7,440,107	10,421,640	2981.53	15	12	2
S114	0125_201	S00_7201_2AA	17	rs858526	rs995362	7,440,107	10,421,640	2981.53	15	14	2
S114	0125_201	S00_7201_2AB	17	rs858526	rs995362	7,440,107	10,421,640	2981.53	15	14	2
S114	0125_201	S00_7201_2AC	17	rs858526	rs995362	7,440,107	10,421,640	2981.53	15	14	2
S114	0125_201	S00_7201_2AD	17	rs858526	rs995362	7,440,107	10,421,640	2981.53	15	14	2
S114	0125_201	00/03743/3A/B	17	rs858526	rs2904912	7,440,107	10,010,470	2570.36	12	12	2
S114	0125_205	98_13905_C	17	rs222836	rs1997095	7,073,886	8,915,247	1841.36	11	6	2
S114	0125_202	99_00083	17	rs1319344	rs1997095	6,113,420	8,915,247	2801.83	15	10	2*
S114	0125_205	04_06591_E	17	rs2309555	rs1997095	6,105,525	8,915,247	2809.72	16	6	2
S114	CON	19	17	rs1565816	rs1997095	7,708,213	8,915,247	1207.03	6	NA	NA
S114	UNION	19	17	rs6502862	rs2240519	1,142,838	11,716,085	10573.2	42	NA	NA
S154	0122_405	M163005_B_2	19	rs887392	rs3499	46,677,642	63,785,296	17107.7	70	15	1
S154	0122_405	M163005_B_3	19	rs887392	rs3499	46,677,642	63,785,296	17107.7	70	15	1
S154	0122_405	M163005_B_6	19	rs887392	rs1673028	46,677,642	55,644,865	8967.22	24	15	1
S154	0122_405	M120505_6_N	19	rs887392	rs741233	46,677,642	51,586,626	4908.98	12	16	1
S154	0122_405	M120505_9_N	19	rs887392	rs741233	46,677,642	51,586,626	4908.98	12	16	1

S154	0122_405	M120505_9_1	19	rs887392	rs741233	46,677,642	51,586,626	4908.98	12	16	1
S154	0122_405	M120505_9_3	19	rs887392	rs741233	46,677,642	51,586,626	4908.98	12	16	1
S154	0122_405	M120505_9_4	19	rs887392	rs741233	46,677,642	51,586,626	4908.98	12	16	1
S154	0122_405	M163005_B_1	19	rs887392	rs741233	46,677,642	51,586,626	4908.98	12	16	1
S154	0122_405	M163005_B_4	19	rs887392	rs741233	46,677,642	51,586,626	4908.98	12	16	1
S154	0122_405	M163005_B_5	19	rs887392	rs741233	46,677,642	51,586,626	4908.98	12	16	1
S154	0122_405	M120505_9_2	19	rs268666	rs1477340	45,610,005	52,221,580	6611.57	17	14	1
S154	0122_301	38990_B	19	rs268666	rs2041975	45,610,005	50,930,395	5320.39	10	15	1
S154	0122_301	38990_A	19	rs268666	rs2041975	45,610,005	50,930,395	5320.39	10	15	1
S154	0122_301	38990_E	19	rs268666	rs2041975	45,610,005	50,930,395	5320.39	10	16	1*
S154	0122_405	M120505_6_T	19	rs575	rs3499	44,124,623	63,785,296	19660.7	78	12	1
S154	0122_304	03/17024_IAA	19	rs1603	rs919364	50,683,476	54,559,725	3876.25	15	12	2
S154	0122_304	07/151131A/A	19	rs993983	rs2041975	44,843,320	50,930,395	6087.07	13	1	2*
S154	0122_304	02/112441A/C	19	rs7937	rs919364	45,994,546	54,559,725	8565.18	22	0	3*
S154	0122_404	06/009346/3A	19	rs887392	rs759623	46,677,642	51,204,357	4526.72	10	0	4*
S154	0122_304	02/112441A/H	19	rs11671074	rs1477340	49,495,959	52,221,580	2725.62	10	0	5*
S154	CON	21	19	rs1603	rs2041975	50,683,476	50,930,395	246.919	2	NA	NA
S154	UNION	21	19	rs575	rs3499	44,124,623	63,785,296	19660.7	78	NA	NA
S47	0065_301	97_11282A2_A	5	rs736201	rs173686	78,873,312	82,847,256	3973.94	10	8	1
S47	0065_301	97_11282A2_B	5	rs736201	rs173686	78,873,312	82,847,256	3973.94	10	8	1
S47	0065_301	97_11282B_A	5	rs736201	rs173686	78,873,312	82,847,256	3973.94	10	8	1
S47	0065_301	97_11282B_B	5	rs736201	rs173686	78,873,312	82,847,256	3973.94	10	8	1
S47	0065_302	97_90571_A	5	rs1200485	rs1020720	72,420,814	80,508,524	8087.71	24	9	1
S47	0065_302	98_41282_A	5	rs1200485	rs34999	72,420,814	80,312,801	7891.99	22	9	1
S47	0065_302	98_41282_B	5	rs1200485	rs34999	72,420,814	80,312,801	7891.99	22	9	1
S47	0065_302	98_41282_C	5	rs1200485	rs34999	72,420,814	80,312,801	7891.99	22	9	1
S47	0065_302	00_8810	5	rs1200485	rs34999	72,420,814	80,312,801	7891.99	22	9	1*

S47	0065_309	A18_0744100_1T	5	rs1531615	rs173686	76,654,613	82,847,256	6192.64	18	6	2
S47	0065_304	9602_1523_3	5	rs463973	rs173686	75,983,781	82,847,256	6863.48	19	1	2*
S47	CON	11	5	rs736201	rs34999	78,873,312	80,312,801	1439.49	3	NA	NA
S47	UNION	11	5	rs1200485	rs173686	72,420,814	82,847,256	10426.4	29	NA	NA
S48	0065_302	98_41282_A	5	rs1554278	rs1820603	37,624,217	40,175,508	2551.29	10	10	1
S48	0065_302	98_41282_B	5	rs1554278	rs1820603	37,624,217	40,175,508	2551.29	10	10	1
S48	0065_302	98_41282_C	5	rs1554278	rs1820603	37,624,217	40,175,508	2551.29	10	10	1
S48	0065_302	00_8810	5	rs1554278	rs1820603	37,624,217	40,175,508	2551.29	10	10	1
S48	0065_309	A18_0744100_1T	5	rs1554278	rs1820603	37,624,217	40,175,508	2551.29	10	10	1
S48	0065_304	9602_1523_3	5	rs2017469	rs476569	31,349,973	39,378,065	8028.09	19	10	1
S48	0065_311	98_1018_A	5	rs2962799	rs476569	31,335,011	39,378,065	8043.05	20	10	1
S48	0065_301	97_11282A2_A	5	rs2034586	rs476569	30,082,244	39,378,065	9295.82	22	10	1
S48	0065_301	97_11282A2_B	5	rs2034586	rs476569	30,082,244	39,378,065	9295.82	22	10	1
S48	0065_301	97_11282B_A	5	rs2034586	rs476569	30,082,244	39,378,065	9295.82	22	10	1
S48	0065_301	97_11282B_B	5	rs2034586	rs476569	30,082,244	39,378,065	9295.82	22	10	1*
S48	CON	11	5	rs1554278	rs476569	37,624,217	39,378,065	1753.85	6	NA	NA
S48	UNION	11	5	rs2034586	rs1820603	30,082,244	40,175,508	10093.3	26	NA	NA
S46	0065_302	00_8810	5	rs270664	rs1363157	158,489,316	163,142,736	4653.42	12	10	1
S46	0065_304	9602_1523_3	5	rs270664	rs1363157	158,489,316	163,142,736	4653.42	12	10	1
S46	0065_309	A18_0744100_1T	5	rs270664	rs1363157	158,489,316	163,142,736	4653.42	12	10	1
S46	0065_302	97_90571_A	5	rs949602	rs1054998	157,323,418	169,548,076	12224.7	27	10	1*
S46	0065_301	97_11282A2_A	5	rs1039322	rs1363157	157,118,638	163,142,736	6024.1	14	7	1
S46	0065_301	97_11282A2_B	5	rs1039322	rs1363157	157,118,638	163,142,736	6024.1	14	7	1
S46	0065_301	97_11282B_A	5	rs1039322	rs1363157	157,118,638	163,142,736	6024.1	14	7	1
S46	0065_301	97_11282B_B	5	rs1039322	rs1363157	157,118,638	163,142,736	6024.1	14	7	1
S46	0065_302	98_41282_B	5	rs385547	rs1053110	122,349,732	180,420,866	58071.1	125	6	1
S46	0065_302	98_41282_A	5	rs3734087	rs1053110	102,922,346	180,420,866	77498.5	159	6	1

S46	0065_302	98_41282_C	5	rs3734087	rs1053110	102,922,346	180,420,866	77498.5	159	6	1
S46	CON	11	5	rs270664	rs1363157	158,489,316	163,142,736	4653.42	12	NA	NA
S46	UNION	11	5	rs3734087	rs1053110	102,922,346	180,420,866	77498.5	159	NA	NA
S38	0088_407	8546_97_3	5	rs1004531	rs1439564	118,632,922	124,587,550	5954.63	16	4	1
S38	0088_522	H004993/B1/A	5	rs27024	rs1004531	97,987,261	118,632,922	20645.7	40	4	1
S38	0088_522	H004993/C1	5	rs1560327	rs1004531	97,053,020	118,632,922	21579.9	43	4	1
S38	0088_522	H004993/A1	5	rs1560327	rs1004531	97,053,020	118,632,922	21579.9	43	4	1
S38	0088_522	H004993/B1/B	5	rs1566629	rs1989154	81,243,028	147,829,083	66586.1	135	4	1*
S38	CON	5	5	rs1004531	rs1004531	118,632,922	118,632,922	0	1	NA	NA
S38	UNION	5	5	rs1566629	rs1989154	81,243,028	147,829,083	66586.1	135	NA	NA
S2	0162_103	99/7772A/A	18	rs1792679	rs917711	43,612,107	52,007,286	8395.18	20	6	1
S2	0162_103	99/7772B/A	18	rs1792679	rs1470325	43,612,107	46,527,289	2915.18	10	6	1*
S2	0162_201	9784/92/2	18	rs920783	rs1145315	42,152,426	49,942,953	7790.53	23	3	1
S2	0162_103	91/1551/1	18	rs8089628	rs1145315	39,722,774	49,942,953	10220.2	27	4	1
S2	0162_203	4958/87	18	rs1878677	rs2456486	38,984,443	58,750,367	19765.9	63	4	1
S2	0162_201	9784/92/1	18	rs1941531	rs652437	38,022,744	52,340,531	14317.8	36	3	1
S2	0162_103	91/1551/2T	18	rs1471408	rs652437	11,531,256	52,340,531	40809.3	89	4	1
S2	CON	7	18	rs1792679	rs1470325	43,612,107	46,527,289	2915.18	10	NA	NA
S2	UNION	7	18	rs1471408	rs2456486	11,531,256	58,750,367	47219.1	117	NA	NA
S187	0122_405	M120505_9_2	18	rs521861	rs13732	45,625,012	53,363,314	7738.3	21	1	1
S187	0122_405	M120505_6_T	18	rs521861	rs1145315	45,625,012	49,942,953	4317.94	10	14	1
S187	0122_405	M120505_6_N	18	rs521861	rs1145315	45,625,012	49,942,953	4317.94	10	14	1
S187	0122_405	M120505_9_N	18	rs521861	rs1145315	45,625,012	49,942,953	4317.94	10	14	1
S187	0122_405	M120505_9_1	18	rs521861	rs1145315	45,625,012	49,942,953	4317.94	10	14	1
S187	0122_405	M120505_9_3	18	rs521861	rs1145315	45,625,012	49,942,953	4317.94	10	14	1
S187	0122_405	M120505_9_4	18	rs521861	rs1145315	45,625,012	49,942,953	4317.94	10	14	1
S187	0122_405	M163005_B_1	18	rs521861	rs1145315	45,625,012	49,942,953	4317.94	10	14	1

S187	0122_405	M163005_B_2	18	rs521861	rs1145315	45,625,012	49,942,953	4317.94	10	14	1
S187	0122_405	M163005_B_3	18	rs521861	rs1145315	45,625,012	49,942,953	4317.94	10	14	1
S187	0122_405	M163005_B_4	18	rs521861	rs1145315	45,625,012	49,942,953	4317.94	10	14	1
S187	0122_405	M163005_B_5	18	rs521861	rs1145315	45,625,012	49,942,953	4317.94	10	14	1
S187	0122_405	M163005_B_6	18	rs521861	rs1145315	45,625,012	49,942,953	4317.94	10	14	1
S187	0122_303	0211284_1A	18	rs7240966	rs1530390	44,649,377	46,569,899	1920.52	10	18	1
S187	0122_303	03_15710_1A	18	rs1981	rs2928927	43,614,989	47,677,183	4062.19	13	6	1
S187	0122_303	01_07226_1A	18	rs1981	rs1822459	43,614,989	46,628,125	3013.14	12	18	1
S187	0122_304	02/112441A/F	18	rs1792679	rs869224	43,612,107	49,231,796	5619.69	15	6	1
S187	0122_304	03/17024_IAA	18	rs1792679	rs1822459	43,612,107	46,628,125	3016.2	13	6	1
S187	0122_304	02/112441A/C	18	rs1502609	rs732982	43,567,388	52,878,457	9311.7	28	6	1
S187	0122_304	07/151131A/A	18	rs1434511	rs1470325	43,083,433	46,527,289	3443.86	13	19	1*
S187	CON	20	18	rs521861	rs1470325	45,625,012	46,527,289	902.277	4	NA	NA
S187	UNION	20	18	rs1434511	rs13732	43,083,433	53,363,314	10279.9	30	NA	NA

Figure 9.7 Genome wide view of LOH for family 336

This figure includes all regions that were greater the 4Mb in size.

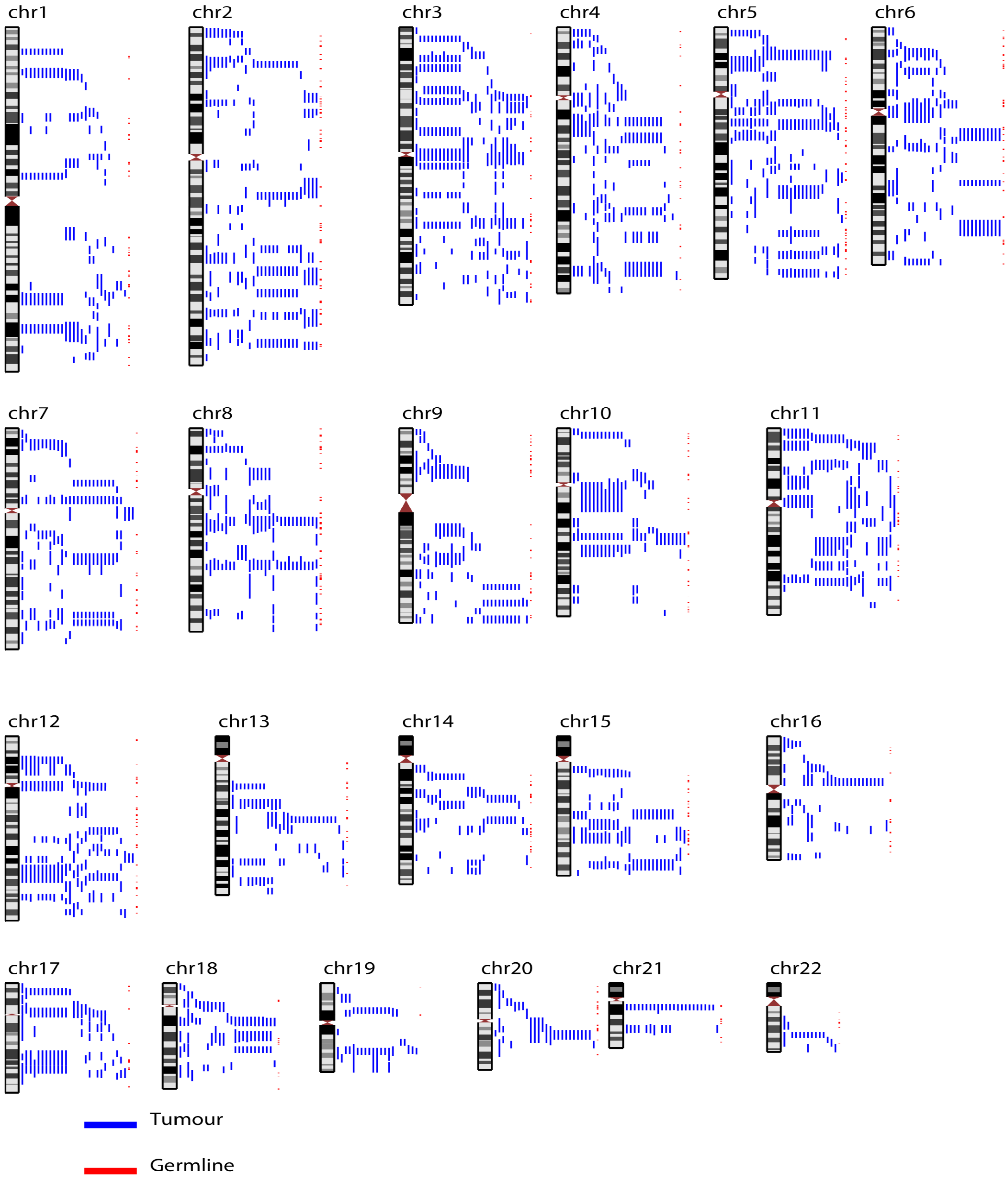


Figure 9.8 Genome-wide view of LOH for family 323

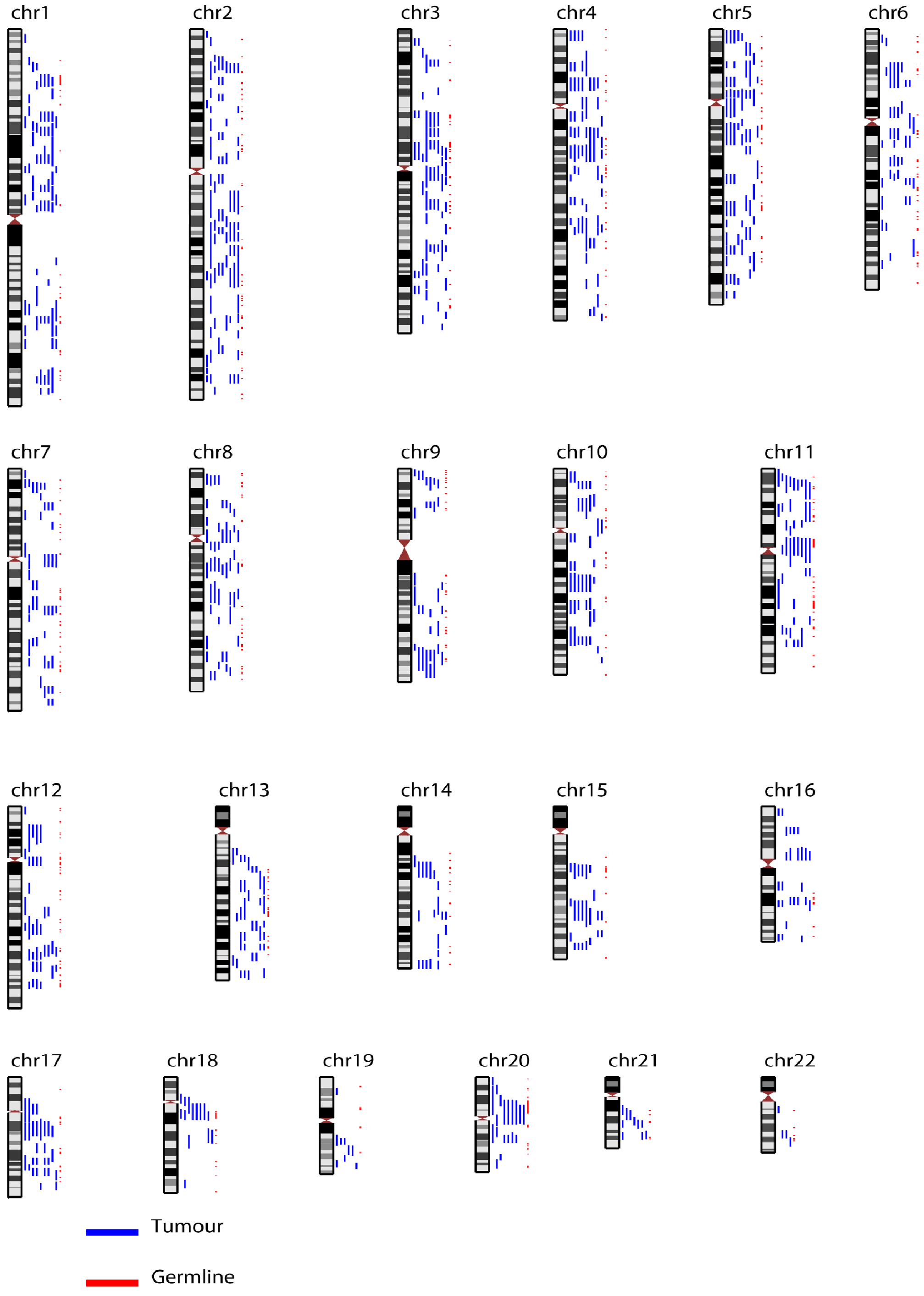


Figure 9.9 Genome-wide view for family 282

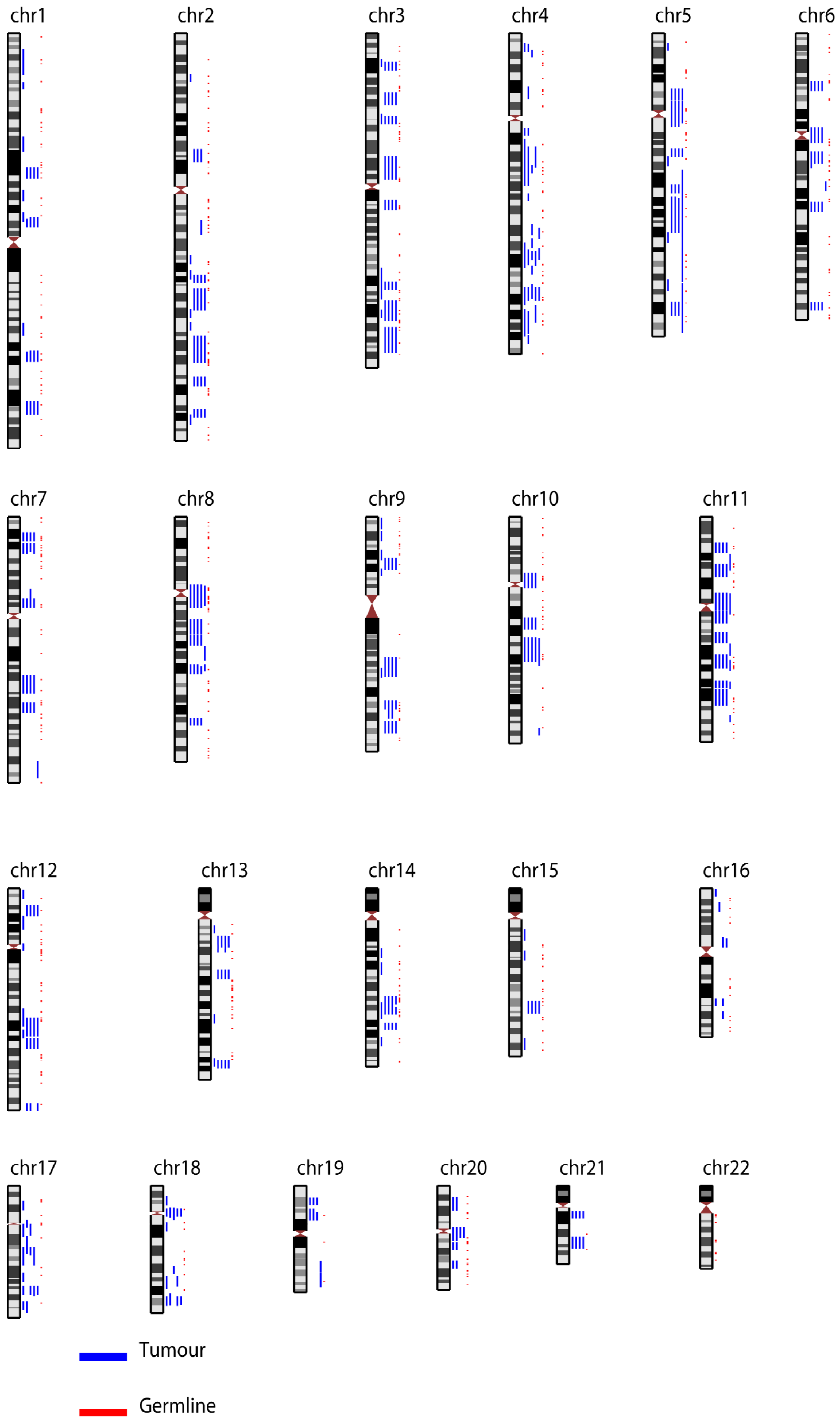


Figure 9.10 Genome-wide view for Family 294

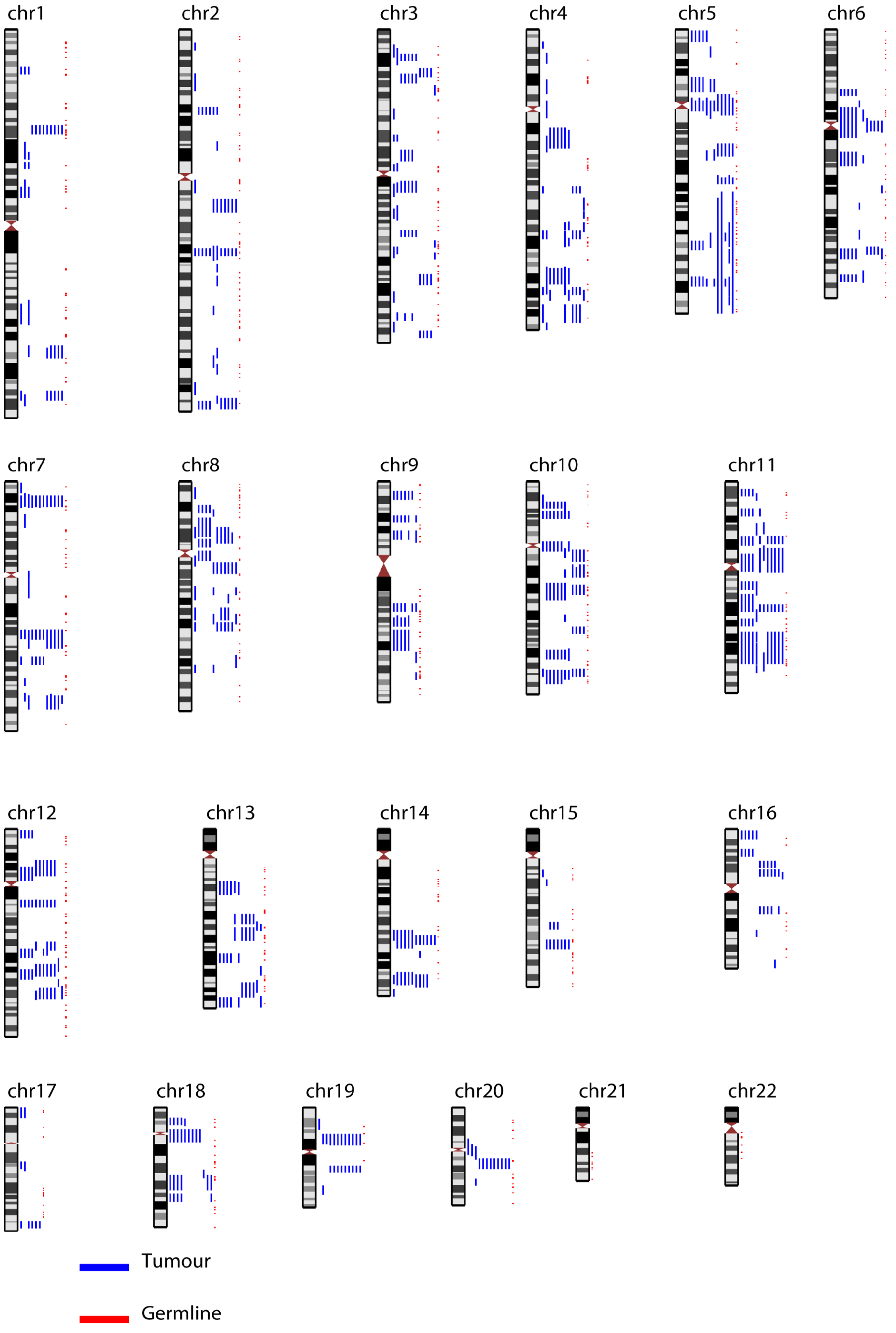


Figure 9.11 Genome-wide view for family 450

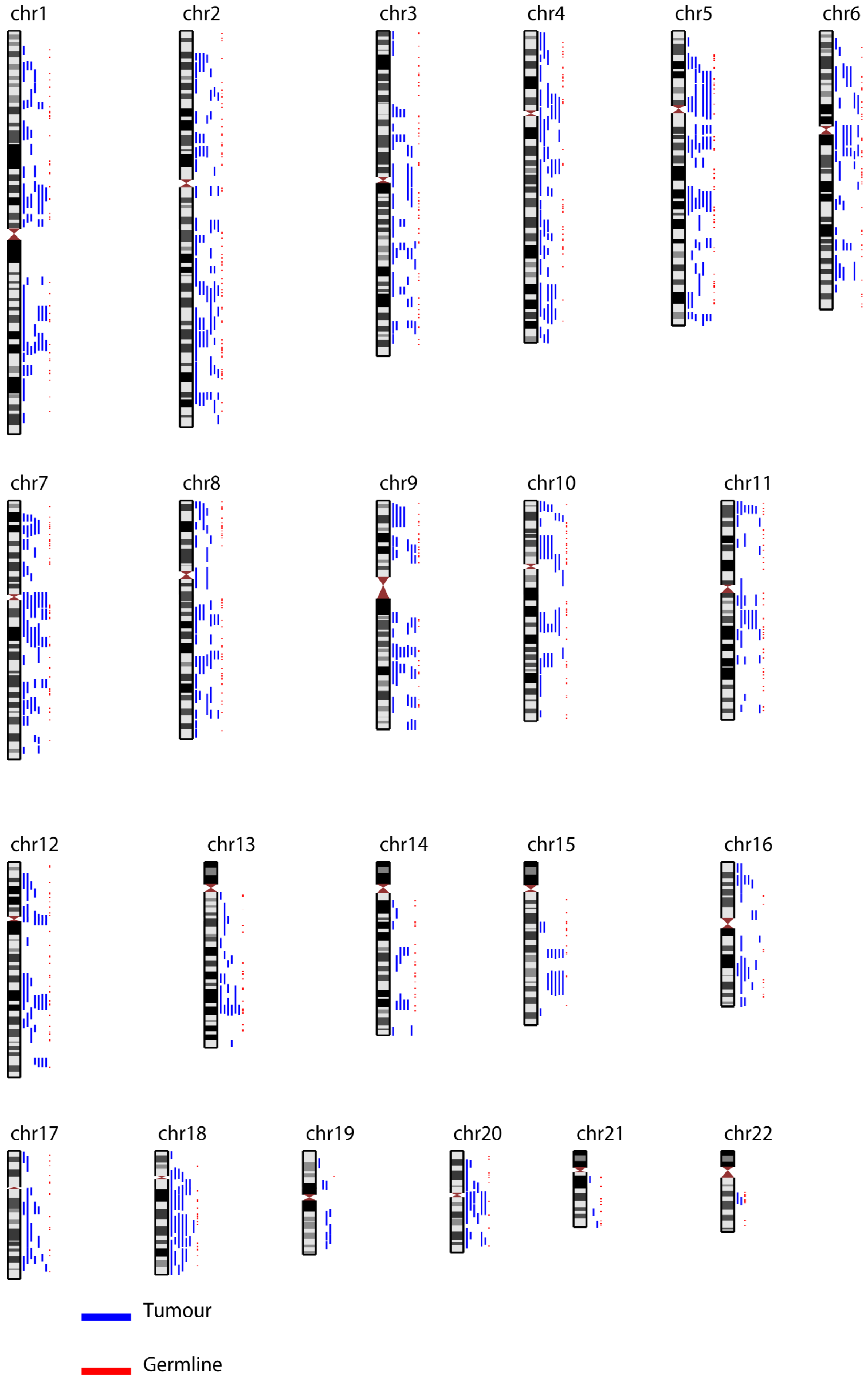


Figure 9.12 Genome-wide view for family 377

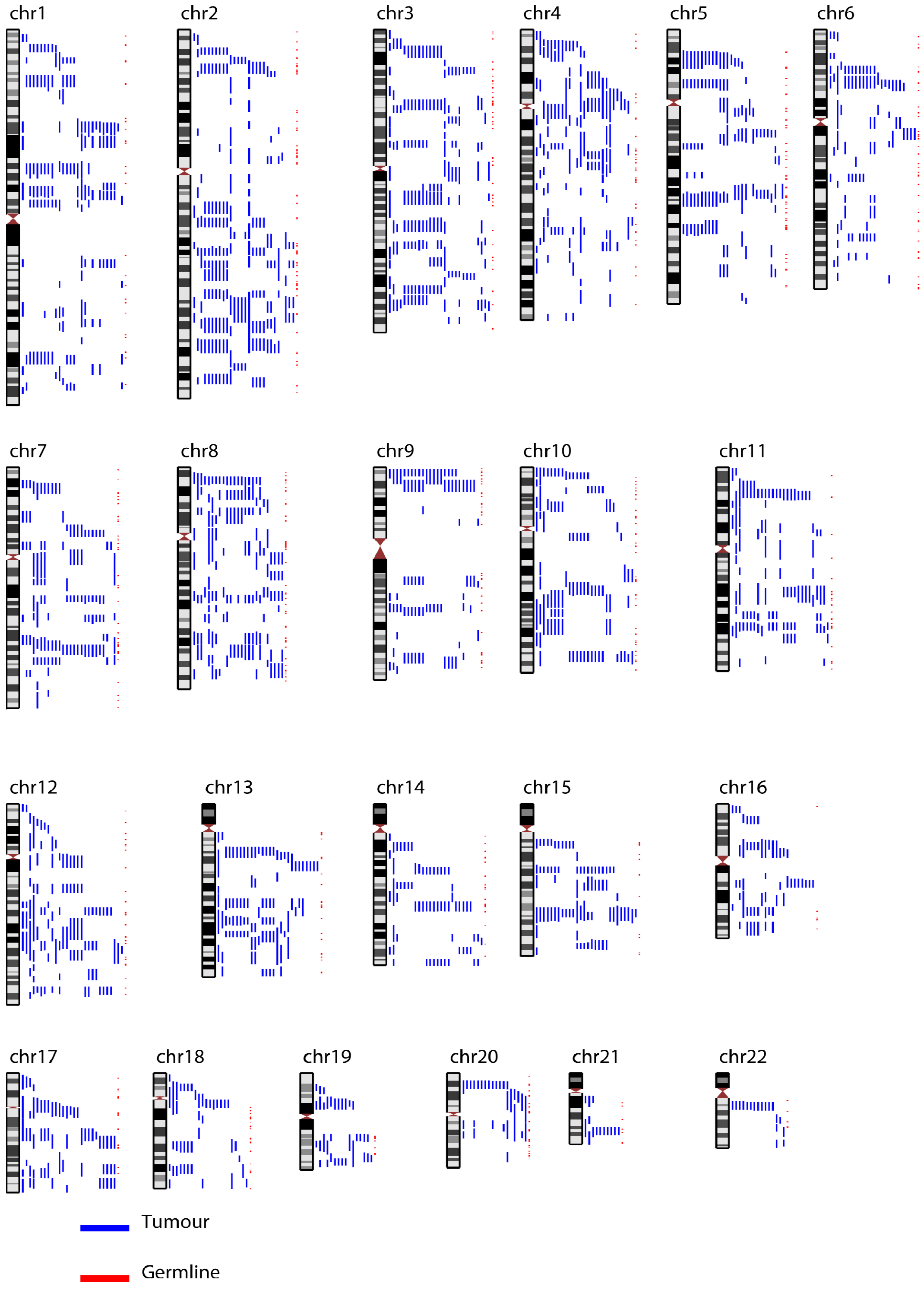


Table 9.4 The individual ROHs that overlap with the detected linkage peaks and contribute to the identified consensus regions listed in Table 7.7

The group column (GRP) indicates the grouping of segments that make up each pool based on segments with 95% allelic identity of the genotypes. A "*" indicates the reference sample for each group.

POOL	FID	IID	CHR	SNP1	SNP2	BP1	BP2	KB	NSNP	NSIM	GRP
S496	0122_303	01_07226_1A	3	rs12634498	rs769276	161,287,222	166,524,940	5237.72	13	14	1
S496	0122_303	03_15710_1A	3	rs12634498	rs769276	161,287,222	166,524,940	5237.72	13	14	1
S496	0122_303	0211284_1A	3	rs6799097	rs769276	160,293,762	166,524,940	6231.18	15	15	1
S496	0122_304	02/112441A/D	3	rs6799097	rs769276	160,293,762	166,524,940	6231.18	15	14	1
S496	0122_304	03/17024_1A	3	rs6799097	rs769276	160,293,762	166,524,940	6231.18	15	14	1
S496	0122_304	02/112441A/F	3	rs6799097	rs769276	160,293,762	166,524,940	6231.18	15	14	1
S496	0122_304	07/151131A/A	3	rs6799097	rs17782339	160,293,762	164,449,715	4155.95	11	15	1*
S496	0122_301	38990_C	3	rs1373118	rs4305435	158,772,638	162,849,335	4076.7	10	8	1
S496	0122_405	M120505_6_N	3	rs1373118	rs4305435	158,772,638	162,849,335	4076.7	10	10	1
S496	0122_405	M120505_9_N	3	rs1373118	rs4305435	158,772,638	162,849,335	4076.7	10	10	1
S496	0122_405	M120505_9_1	3	rs1373118	rs4305435	158,772,638	162,849,335	4076.7	10	10	1
S496	0122_405	M163005_B_4	3	rs1373118	rs4305435	158,772,638	162,849,335	4076.7	10	10	1
S496	0122_301	38990_B	3	rs1074864	rs1492174	158,615,279	164,450,385	5835.11	16	3	1
S496	0122_304	02/112441A/C	3	rs9438	rs17782339	155,501,589	164,449,715	8948.13	23	9	1
S496	0122_404	06/009346/3A	3	rs755763	rs769276	153,482,627	166,524,940	13042.3	31	9	1
S496	0122_403	98/685_1A	3	rs755763	rs769276	153,482,627	166,524,940	13042.3	31	9	1
S496	CON	16	3	rs12634498	rs4305435	161,287,222	162,849,335	1562.11	5	NA	NA
S496	UNION	16	3	rs755763	rs769276	153,482,627	166,524,940	13042.3	31	NA	NA
S1119	0122_303	02/11284/2A/D	3	rs1450344	rs359573	151,668,443	156,803,536	5135.09	13	3	1
S1119	0122_303	01_07226_1A	3	rs1561026	rs359573	150,834,330	156,803,536	5969.21	14	2	1
S1119	0122_303	03_15710_1A	3	rs1561026	rs359573	150,834,330	156,803,536	5969.21	14	2	1
S1119	0122_304	02/112441A/I	3	rs1398775	rs6773566	147,030,054	156,507,016	9476.96	23	5	1*

S1119	0122_404	06/009346/3A	3	rs755763	rs769276	153,482,627	166,524,940	13042.3	31	5	2
S1119	0122_403	98/685_1A	3	rs755763	rs769276	153,482,627	166,524,940	13042.3	31	5	2
S1119	0122_405	M120505_6_T	3	rs1009621	rs1388007	149,673,446	154,738,672	5065.23	10	3	2
S1119	0122_405	M163005_B_2	3	rs1009621	rs1388007	149,673,446	154,738,672	5065.23	10	3	2*
S1119	CON	8	3	rs755763	rs1388007	153,482,627	154,738,672	1256.05	3	NA	NA
S1119	UNION	8	3	rs1398775	rs769276	147,030,054	166,524,940	19494.9	46	NA	NA
S1232	0122_304	02/112441A/C	3	rs9438	rs17782339	155,501,589	164,449,715	8948.13	23	6	1
S1232	0122_404	06/009346/3A	3	rs755763	rs769276	153,482,627	166,524,940	13042.3	31	4	1
S1232	0122_403	98/685_1A	3	rs755763	rs769276	153,482,627	166,524,940	13042.3	31	4	1
S1232	0122_303	02/11284/2A/D	3	rs1450344	rs359573	151,668,443	156,803,536	5135.09	13	4	1
S1232	0122_303	01_07226_1A	3	rs1561026	rs359573	150,834,330	156,803,536	5969.21	14	3	1
S1232	0122_303	03_15710_1A	3	rs1561026	rs359573	150,834,330	156,803,536	5969.21	14	3	1
S1232	0122_304	02/112441A/I	3	rs1398775	rs6773566	147,030,054	156,507,016	9476.96	23	6	1*
S1232	CON	7	3	rs9438	rs6773566	155,501,589	156,507,016	1005.43	4	NA	NA
S1232	UNION	7	3	rs1398775	rs769276	147,030,054	166,524,940	19494.9	46	NA	NA
S350	0125_204	97/13626/2A	3	rs1799404	rs1919987	127,641,320	133,159,282	5517.96	16	14	1
S350	0125_205	04_06591_E	3	rs13975	rs1402455	126,284,921	134,196,823	7911.9	21	14	1
S350	0125_205	98_13905_C	3	rs986909	rs719300	125,766,591	133,977,145	8210.55	21	14	1
S350	0125_204	97_13626_1A	3	rs986909	rs719300	125,766,591	133,977,145	8210.55	21	14	1
S350	0125_201	S0110291/1A/A	3	rs986909	rs2199351	125,766,591	133,090,237	7323.65	19	14	1
S350	0125_201	00/03743/3A/A	3	rs986909	rs6792114	125,766,591	133,018,080	7251.49	18	14	1
S350	0125_201	00/03743/3A/B	3	rs986909	rs6792114	125,766,591	133,018,080	7251.49	18	14	1
S350	0125_201	00/03743/3A/C	3	rs986909	rs6792114	125,766,591	133,018,080	7251.49	18	14	1
S350	0125_201	00/03743/1A	3	rs986909	rs6792114	125,766,591	133,018,080	7251.49	18	14	1
S350	0125_201	S0110291/1A/C	3	rs986909	rs6792114	125,766,591	133,018,080	7251.49	18	14	1
S350	0125_201	S0110291/1A/D	3	rs986909	rs6792114	125,766,591	133,018,080	7251.49	18	14	1
S350	0125_201	S00_7201_2AA	3	rs1127343	rs2199351	123,611,084	133,090,237	9479.15	25	14	1

S350	0125_201	S00_7201_2AB	3	rs1127343	rs2199351	123,611,084	133,090,237	9479.15	25	14	1
S350	0125_201	S00_7201_2AC	3	rs1127343	rs2199351	123,611,084	133,090,237	9479.15	25	14	1
S350	0125_201	S00_7201_2AD	3	rs1127343	rs2199351	123,611,084	133,090,237	9479.15	25	14	1*
S350	CON	15	3	rs1799404	rs6792114	127,641,320	133,018,080	5376.76	14	NA	NA
S350	UNION	15	3	rs1127343	rs1402455	123,611,084	134,196,823	10585.7	28	NA	NA
S1524	0125_201	S00_7201_2AA	3	rs1127343	rs2199351	123,611,084	133,090,237	9479.15	25	3	1
S1524	0125_201	S00_7201_2AB	3	rs1127343	rs2199351	123,611,084	133,090,237	9479.15	25	3	1
S1524	0125_201	S00_7201_2AC	3	rs1127343	rs2199351	123,611,084	133,090,237	9479.15	25	3	1
S1524	0125_201	S00_7201_2AD	3	rs1127343	rs2199351	123,611,084	133,090,237	9479.15	25	3	1*
S1524	0125_302	044123_1A_C	3	rs1147696	rs634265	121,602,169	125,586,787	3984.62	10	0	2*
S1524	CON	5	3	rs1127343	rs634265	123,611,084	125,586,787	1975.7	5	NA	NA
S1524	UNION	5	3	rs1147696	rs2199351	121,602,169	133,090,237	11488.1	30	NA	NA
S2011	0125_302	044123_1A_C	3	rs1147696	rs634265	121,602,169	125,586,787	3984.62	10	1	1
S2011	0125_302	044123_1A_B	3	rs1436340	rs1472621	106,080,221	123,495,416	17415.2	37	1	1*
S2011	CON	2	3	rs1147696	rs1472621	121,602,169	123,495,416	1893.25	5	NA	NA
S2011	UNION	2	3	rs1436340	rs634265	106,080,221	125,586,787	19506.6	42	NA	NA
S16	0065_301	97_11282A2_A	7	rs6463843	rs1723804	8,611,957	16,557,184	7945.23	17	11	1
S16	0065_301	97_11282B_A	7	rs6463843	rs1723804	8,611,957	16,557,184	7945.23	17	11	1
S16	0065_304	9602_1523_3	7	rs6463843	rs1723804	8,611,957	16,557,184	7945.23	17	11	1
S16	0065_301	97_11282A2_B	7	rs6463843	rs2030972	8,611,957	16,025,705	7413.75	16	11	1
S16	0065_301	97_11282B_B	7	rs6463843	rs2030972	8,611,957	16,025,705	7413.75	16	11	1*
S16	0065_302	98_41282_A	7	rs6463843	rs2030972	8,611,957	16,025,705	7413.75	16	9	1
S16	0065_302	98_41282_C	7	rs6463843	rs2030972	8,611,957	16,025,705	7413.75	16	9	1
S16	0065_302	97_90571_A	7	rs6463843	rs2030972	8,611,957	16,025,705	7413.75	16	9	1
S16	0065_302	98_41282_B	7	rs6463843	rs1029718	8,611,957	15,817,540	7205.58	15	9	1
S16	0065_302	00_8810	7	rs6463843	rs1029718	8,611,957	15,817,540	7205.58	15	9	1
S16	0065_311	A18_0744100_1T	7	rs37995	rs1723804	7,802,374	16,557,184	8754.81	19	6	1

S16	0065_302	98_1018_A	7	rs13438509	rs1029718	7,142,832	15,817,540	8674.71	19	6	1
S16	CON	12	7	rs6463843	rs1029718	8,611,957	15,817,540	7205.58	15	NA	NA
S16	UNION	12	7	rs13438509	rs1723804	7,142,832	16,557,184	9414.35	21	NA	NA
S1626	0125_201	S0110291/1A/D	9	rs1128957	rs17179086	25,667,257	30,897,980	5230.72	12	3	1
S1626	0125_201	00/03743/3A/A	9	rs1128957	rs560764	25,667,257	29,582,836	3915.58	10	3	1
S1626	0125_204	97/13626/2A	9	rs1128957	rs560764	25,667,257	29,582,836	3915.58	10	3	1
S1626	0125_204	97/13626/3A	9	rs1128957	rs560764	25,667,257	29,582,836	3915.58	10	3	1*
S1626	CON	4	9	rs1128957	rs560764	25,667,257	29,582,836	3915.58	10	NA	NA
S1626	UNION	4	9	rs1128957	rs17179086	25,667,257	30,897,980	5230.72	12	NA	NA
S217	0125_302	044123_1A_D	9	rs7866589	rs263580	14,237,788	17,029,312	2791.52	10	16	1
S217	0125_203	95_2882_C_T	9	rs1156793	rs1328273	8,368,662	16,013,469	7644.81	14	16	1
S217	0125_203	95_2882_C_N	9	rs1156793	rs1328273	8,368,662	16,013,469	7644.81	14	16	1
S217	0125_205	98_13905_C	9	rs1156793	rs1328273	8,368,662	16,013,469	7644.81	14	16	1
S217	0125_204	97_13626_1A	9	rs1156793	rs1328273	8,368,662	16,013,469	7644.81	14	16	1
S217	0125_201	00/03743/3A/A	9	rs1156793	rs1328273	8,368,662	16,013,469	7644.81	14	16	1
S217	0125_201	00/03743/3A/B	9	rs1156793	rs1328273	8,368,662	16,013,469	7644.81	14	16	1
S217	0125_201	00/03743/3A/C	9	rs1156793	rs1328273	8,368,662	16,013,469	7644.81	14	16	1
S217	0125_201	00/03743/1A	9	rs1156793	rs1328273	8,368,662	16,013,469	7644.81	14	16	1
S217	0125_201	S0110291/1A/A	9	rs1156793	rs1328273	8,368,662	16,013,469	7644.81	14	16	1
S217	0125_201	S0110291/1A/D	9	rs1156793	rs1328273	8,368,662	16,013,469	7644.81	14	16	1
S217	0125_204	97/13626/2A	9	rs1156793	rs1328273	8,368,662	16,013,469	7644.81	14	16	1
S217	0125_204	97/13626/3A	9	rs1156793	rs1328273	8,368,662	16,013,469	7644.81	14	16	1
S217	0125_201	S00_7201_2AA	9	rs1156793	rs1328273	8,368,662	16,013,469	7644.81	14	16	1
S217	0125_201	S00_7201_2AB	9	rs1156793	rs1328273	8,368,662	16,013,469	7644.81	14	16	1
S217	0125_201	S00_7201_2AC	9	rs1156793	rs1328273	8,368,662	16,013,469	7644.81	14	16	1
S217	0125_201	S00_7201_2AD	9	rs1156793	rs1328273	8,368,662	16,013,469	7644.81	14	16	1*
S217	CON	17	9	rs7866589	rs1328273	14,237,788	16,013,469	1775.68	5	NA	NA

S217	UNION	17	9	rs1156793	rs263580	8,368,662	17,029,312	8660.65	19	NA	NA
S1663	0122_301	38990_C	10	rs911610	rs1227938	64,290,048	70,828,274	6538.23	10	3	1
S1663	0122_301	38990_B	10	rs911610	rs1227938	64,290,048	70,828,274	6538.23	10	3	1
S1663	0122_301	38990_A	10	rs911610	rs1227938	64,290,048	70,828,274	6538.23	10	3	1
S1663	0122_301	38990_E	10	rs911610	rs1227938	64,290,048	70,828,274	6538.23	10	3	1*
S1663	CON	4	10	rs911610	rs1227938	64,290,048	70,828,274	6538.23	10	NA	NA
S1663	UNION	4	10	rs911610	rs1227938	64,290,048	70,828,274	6538.23	10	NA	NA
S146	0122_403	98/685_1A	2	rs1431087	rs475525	226,309,338	230,559,586	4250.25	10	21	1
S146	0122_301	38990_C	2	rs375154	rs730010	225,705,579	231,004,381	5298.8	15	17	1
S146	0122_301	38990_B	2	rs375154	rs730010	225,705,579	231,004,381	5298.8	15	20	1
S146	0122_301	38990_A	2	rs375154	rs730010	225,705,579	231,004,381	5298.8	15	17	1
S146	0122_301	38990_E	2	rs375154	rs730010	225,705,579	231,004,381	5298.8	15	17	1
S146	0122_405	M120505_6_T	2	rs959327	rs2396611	223,436,304	229,652,416	6216.11	21	21	1
S146	0122_405	M120505_6_N	2	rs959327	rs2396611	223,436,304	229,652,416	6216.11	21	21	1
S146	0122_405	M120505_9_N	2	rs959327	rs2396611	223,436,304	229,652,416	6216.11	21	21	1
S146	0122_405	M120505_9_1	2	rs959327	rs2396611	223,436,304	229,652,416	6216.11	21	21	1
S146	0122_405	M120505_9_2	2	rs959327	rs2396611	223,436,304	229,652,416	6216.11	21	21	1
S146	0122_405	M120505_9_3	2	rs959327	rs2396611	223,436,304	229,652,416	6216.11	21	21	1
S146	0122_405	M120505_9_4	2	rs959327	rs2396611	223,436,304	229,652,416	6216.11	21	21	1
S146	0122_405	M163005_B_1	2	rs959327	rs2396611	223,436,304	229,652,416	6216.11	21	21	1
S146	0122_405	M163005_B_2	2	rs959327	rs2396611	223,436,304	229,652,416	6216.11	21	21	1
S146	0122_405	M163005_B_3	2	rs959327	rs2396611	223,436,304	229,652,416	6216.11	21	21	1
S146	0122_405	M163005_B_4	2	rs959327	rs2396611	223,436,304	229,652,416	6216.11	21	21	1
S146	0122_405	M163005_B_5	2	rs959327	rs2396611	223,436,304	229,652,416	6216.11	21	21	1
S146	0122_405	M163005_B_6	2	rs959327	rs2396611	223,436,304	229,652,416	6216.11	21	21	1*
S146	0122_303	0211284_1A	2	rs959327	rs997363	223,436,304	228,334,963	4898.66	18	18	1
S146	0122_304	02/112441A/D	2	rs348971	rs997363	222,743,696	228,334,963	5591.27	20	18	1

S146	0122_304	02/112441A/I	2	rs1554622	rs730010	219,431,723	231,004,381	11572.7	33	17	1
S146	0122_304	02/112441A/C	2	rs207928	rs997363	216,861,947	228,334,963	11473	32	18	1
S146	CON	22	2	rs1431087	rs997363	226,309,338	228,334,963	2025.62	4	NA	NA
S146	UNION	22	2	rs207928	rs730010	216,861,947	231,004,381	14142.4	39	NA	NA
S11	0122_301	38990_C	2	rs375154	rs730010	225,705,579	231,004,381	5298.8	15	24	1
S11	0122_301	38990_B	2	rs375154	rs730010	225,705,579	231,004,381	5298.8	15	27	1
S11	0122_301	38990_A	2	rs375154	rs730010	225,705,579	231,004,381	5298.8	15	24	1
S11	0122_301	38990_E	2	rs375154	rs730010	225,705,579	231,004,381	5298.8	15	24	1
S11	0122_405	M120505_6_T	2	rs959327	rs2396611	223,436,304	229,652,416	6216.11	21	28	1
S11	0122_405	M120505_6_N	2	rs959327	rs2396611	223,436,304	229,652,416	6216.11	21	28	1
S11	0122_405	M120505_9_N	2	rs959327	rs2396611	223,436,304	229,652,416	6216.11	21	28	1
S11	0122_405	M120505_9_1	2	rs959327	rs2396611	223,436,304	229,652,416	6216.11	21	28	1
S11	0122_405	M120505_9_2	2	rs959327	rs2396611	223,436,304	229,652,416	6216.11	21	28	1
S11	0122_405	M120505_9_3	2	rs959327	rs2396611	223,436,304	229,652,416	6216.11	21	28	1
S11	0122_405	M120505_9_4	2	rs959327	rs2396611	223,436,304	229,652,416	6216.11	21	28	1
S11	0122_405	M163005_B_1	2	rs959327	rs2396611	223,436,304	229,652,416	6216.11	21	28	1
S11	0122_405	M163005_B_2	2	rs959327	rs2396611	223,436,304	229,652,416	6216.11	21	28	1
S11	0122_405	M163005_B_3	2	rs959327	rs2396611	223,436,304	229,652,416	6216.11	21	28	1
S11	0122_405	M163005_B_4	2	rs959327	rs2396611	223,436,304	229,652,416	6216.11	21	28	1
S11	0122_405	M163005_B_5	2	rs959327	rs2396611	223,436,304	229,652,416	6216.11	21	28	1
S11	0122_405	M163005_B_6	2	rs959327	rs2396611	223,436,304	229,652,416	6216.11	21	28	1
S11	0122_303	0211284_1A	2	rs959327	rs997363	223,436,304	228,334,963	4898.66	18	25	1
S11	0122_304	03/17024_1A	2	rs959327	rs1431079	223,436,304	226,240,970	2804.67	14	28	1
S11	0122_304	02/112441A/F	2	rs959327	rs1431079	223,436,304	226,240,970	2804.67	14	28	1
S11	0122_304	02/112441A/H	2	rs959327	rs1431079	223,436,304	226,240,970	2804.67	14	28	1
S11	0122_303	02/11284/2A/D	2	rs959327	rs936070	223,436,304	226,209,380	2773.08	13	28	1
S11	0122_304	02/112441A/D	2	rs348971	rs997363	222,743,696	228,334,963	5591.27	20	25	1

S11	0122_304	07/151131A/A	2	rs348971	rs1431079	222,743,696	226,240,970	3497.27	16	28	1
S11	0122_304	03/17024_IAA	2	rs348971	rs1431079	222,743,696	226,240,970	3497.27	16	28	1*
S11	0122_304	02/112441A/I	2	rs1554622	rs730010	219,431,723	231,004,381	11572.7	33	24	1
S11	0122_303	01_07226_1A	2	rs1851328	rs1431079	217,006,038	226,240,970	9234.93	27	27	1
S11	0122_303	03_15710_1A	2	rs1851328	rs1431079	217,006,038	226,240,970	9234.93	27	27	1
S11	0122_304	02/112441A/C	2	rs207928	rs997363	216,861,947	228,334,963	11473	32	23	1
S11	CON	29	2	rs375154	rs936070	225,705,579	226,209,380	503.801	3	NA	NA
S11	UNION	29	2	rs207928	rs730010	216,861,947	231,004,381	14142.4	39	NA	NA
S228	0125_302	044123_1A_B	2	rs896441	rs869134	195,091,738	201,384,483	6292.74	10	16	1
S228	0125_201	S00_7201_2AA	2	rs1882395	rs1455335	191,510,509	199,368,224	7857.72	10	15	1
S228	0125_201	S00_7201_2AB	2	rs1882395	rs1455335	191,510,509	199,368,224	7857.72	10	15	1
S228	0125_201	S00_7201_2AC	2	rs1882395	rs1455335	191,510,509	199,368,224	7857.72	10	15	1
S228	0125_201	S00_7201_2AD	2	rs1882395	rs1455335	191,510,509	199,368,224	7857.72	10	15	1
S228	0125_205	98_13905_C	2	rs3791767	rs7014	190,348,160	196,350,910	6002.75	10	16	1
S228	0125_204	97_13626_1A	2	rs3791767	rs7014	190,348,160	196,350,910	6002.75	10	16	1
S228	0125_204	97/13626/2A	2	rs3791767	rs7014	190,348,160	196,350,910	6002.75	10	16	1
S228	0125_201	00/03743/3A/A	2	rs3134656	rs1455335	189,564,916	199,368,224	9803.31	15	15	1
S228	0125_201	00/03743/3A/B	2	rs3134656	rs1455335	189,564,916	199,368,224	9803.31	15	15	1
S228	0125_201	00/03743/3A/C	2	rs3134656	rs1455335	189,564,916	199,368,224	9803.31	15	15	1
S228	0125_201	00/03743/1A	2	rs3134656	rs1455335	189,564,916	199,368,224	9803.31	15	15	1
S228	0125_201	S0110291/1A/A	2	rs3134656	rs1455335	189,564,916	199,368,224	9803.31	15	15	1
S228	0125_201	S0110291/1A/C	2	rs3134656	rs1455335	189,564,916	199,368,224	9803.31	15	15	1
S228	0125_201	S0110291/1A/D	2	rs3134656	rs1455335	189,564,916	199,368,224	9803.31	15	15	1
S228	0125_204	97/13626/3A	2	rs1354905	rs1455335	189,271,977	199,368,224	10096.2	16	5	1
S228	0125_205	04_06591_E	2	rs925881	rs1369842	172,529,148	200,934,493	28405.3	47	16	1*
S228	CON	17	2	rs896441	rs7014	195,091,738	196,350,910	1259.17	3	NA	NA
S228	UNION	17	2	rs925881	rs869134	172,529,148	201,384,483	28855.3	50	NA	NA

S2019	0125_302	044123_1A_B	2	rs896441	rs869134	195,091,738	201,384,483	6292.74	10	0	1*
S2019	0125_205	04_06591_E	2	rs2715896	rs714393	201,266,795	212,524,224	11257.4	25	0	2*
S2019	CON	2	2	rs2715896	rs869134	201,266,795	201,384,483	117.688	2	NA	NA
S2019	UNION	2	2	rs896441	rs714393	195,091,738	212,524,224	17432.5	33	NA	NA