Germline determinants of outcome and risk in colorectal cancer

Axel Walther

Thesis submitted for the PhD examination at University College London

To Paul and Sophie

I, Axel Walther, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, this has been indicated in the thesis. Parts of the introduction, in particular section 1.6, have been published as a review during the writing of this thesis[1] and received some input from the co-authors of that review.

London, 1 February 2010

Abstract

Genome-wide association studies (GWAS) have identified germline single nucleotide polymorphisms (SNPs) that are associated with colorectal cancer (CRC) susceptibility. This thesis applies the same approach to the identification of germline determinants of prognosis in CRC, attempts to verify potential susceptibility loci, and examines the relationship between SNPs and some forms of non-SNP based germline variation.

The GWAS for prognosis used 931 patients enrolled in the VICTOR trial in the discovery phase, screening 309,200 autosomal SNPs for an association with disease-free survival (DFS). Following the application of selection filters based on statistical significance levels and performance of the genotyping, 40 SNPs were identified to be examined in further cohorts. The verification phase consisted of 1338 patients in the PETACC 3 trial and three population based cohorts: 899 patients from Scotland, 599 patients from Denmark, and 962 patients from Finland.

The SNPs that came closest to genome-wide significance in stage 2 and 3 CRC was rs7556894, 15kb from Actin-related protein 2 (*ARP2*) on chromosome 2, part of the ARP2/3 complex essential for cell shape and motility, with p=8.96e-07. The impact on prognosis of rs7556894 was estimated as HR=1.52 (95% CI 1.17-1.96).

Because of the failure to reach genome-wide significance (p<1e-07), two further approaches to the discovery phase are presented: the meta-analysis of two discovery cohorts to increase event rate and subject numbers and a GWAS for predictive markers for the benefit of adjuvant 5-FU chemotherapy. Formal verification of either approach was not undertaken as part of this thesis.

Further loci were subjected to specific analyses of association with prognosis or CRC susceptibility: rs6983267 and the previously identified CRC susceptibility loci to a survival analysis, and not found to be associated; rs6687758, previously identified as a potential CRC risk locus to a susceptibility verification, confirming a significant association with HR=1.15, 95% CI 1.10-1.21, p=5.04e-08; and a variety of hypothesis driven potential risk loci to a screen for an association with CRC susceptibility, none was found but the LD relationship between tagSNPs and insertion/deletion polymorphisms appears to be the same as for 'normal' SNPs.

Overall, the data presented in this thesis quantify further the contribution of germline variation to CRC susceptibility, exclude a major effect of such variation on prognosis, and verify rs6687758 as a further low-penetrance CRC susceptibility locus.

Table of Figures

Acknowledgements

I would like to express my gratitude to my supervisor, Prof Ian Tomlinson, for the opportunities afforded to me in completing this PhD, for his support and intellectual stimulation. My sincere thanks also go to Prof David Kerr for his support and for allowing me to use the samples from the VICTOR study as basis for this thesis.

I would like to thank everybody in the Molecular and Population Genetics laboratory for their support and for creating a fun and challenging working environment, in particular Andrew Rowan for putting up with all my questions on how to do 'things', Dr Enric Domingo for helping with the genotyping of the VICTOR samples, Sarah Spain, Kimberley Howarth, Dr Luis Cavarjal-Camona, and Dr Jean-Baptiste Cazier for sharing genotype data and discussions, and the latter also for help with the PCA analysis of VICTOR, and Angela Jones and Dr Emma Jaeger for answering yet more questions. I would like to thank everybody in the Equipment Park at the London Research Institute for their help, and Dr Charles Swanton for the scientific discussions.

Thank you also to Dr Elaine Johnstone, Haitao Wang and Patrick Julier in Oxford for helping with DNA and data for the VICTOR samples.

I would like to thank David Mesher and Prof Peter Sasieni at the Wolfson Institute of Preventive Medicine for their support and patience in matters statistical, Prof Sabine Tejpar and Dr Mauro Delorenzi and their colleagues at the EORTC for allowing me access to genotypes and survival statistics for the PETACC 3 study, Prof Malcolm Dunlop and Dr Albert Tenesa in Edinburgh for sharing the genotypes and survival data for the Scottish phase 1, and all collaborators listed below who provided samples, data, and support.

Finally, I would like to thank my wife Nicola for her patience, love, and childcare during the long hours while I completed this thesis.

Institution	Collaborators
University Hospital Gasthuisberg, Leuven, Belgium	Bart Biesmans, Dr Diether Lambrechts
Geneva University Hospital, Switzerland	Dr Arnaud Roth
Department of Molecular Medicine, Aarhus University, Denmark	Dr Claus Lindbjerg Andersen, Prof Torben Ørntoft
Department of Medical Genetics, Biomedicum Helsinki, Finland	Dr Sari Tuupanen, Prof Lauri Aaltonen
Institute for Cancer Research, Sutton	Prof Richard Houlston
Ludwig Institute for Cancer Research, Melbourne, Australia	Dr Oliver Sieber, Dr Lara Lipton
Hospital Clínic, Barcelona, Spain	Dr Serge Castellvi-Bel
Health Sciences Research Institute, Warwick University	Christopher McConkey

Abbreviations

bp	base pair
CI	confidence interval
CIMP	CpG island methylator phenotype
CIN	chromosomal instability
CNV	copy number variation
CRC	colorectal cancer
df	degrees of freedom
ES	effect size (either HR or OR)
EST	expressed sequence tag
5-FU	5-fluorouracil
FFPE	formalin-fixed, paraffin-embedded
FAP	familial adenomatous polyposis
GRR	genotypic relative risk
GWAS	genome-wide association study
Hap300 arrays	Illumina HumanHap300 Duo BeadChips
Hap370 arrays	Illumina HumanHap370 Duo BeadChips
HNPCC	hereditary non-polyposis colorectal cancer
HR	hazard ratio
HWE	Hardy-Weinberg equilibrium
JPS	juvenile polyposis syndrome
Kb	kilo base pairs (10^3 base pairs)
LCI	lower confidence interval
LD	linkage disequilibrium
MAF	minor allele frequency
Mb	mega base pairs (10^6 base pairs)
MMR genes	mismatch repair genes
MSI	microsatellite instability
NGS	next generation sequencing
OR	odds radio
ORF	open reading frame
PJS	Peutz-Jeghers syndrome
p_{all}	p-value for allelic model
p_{dom}	p-value for dominant model
p_{gen}	p-value for genotypic model
p_{het}	p-vale for comparison of heterozygote groups
p_{hom}	p-value for comparison of homozygote groups
p_{Q}	p-value associated with heterogeneity
p_{rec}	p-value for recessive model
SD	standard deviation
SNP	single nucleotide polymorphism
Stage	AJCC stage
TNM	tumour, lymphnode, metastasis stage
TS	thymidylate synthetase
UCI	upper confidence interval

Definitions and Conventions

BeadChip Duo

The Illumina SNP array, so-called because hundreds of thousands of Beadtypes are hybridised to the array, and because each array can be used to analyse two samples.

BeadStudio

The Illumina genotype analysis software.

Beadtype

A bead with several multimer DNA sequences attached. Each multimer consists of 50 bases for target sequence recognition (the SNP), and 22 bases for determining positional information on the array (the IllumiCode). Each bead carries tags for only one SNP, and each Beadtype is replicated approximately 15-17 times per Duo array

Call frequency

A SNP score giving the percentage of samples called for that SNP.

Call rate

A sample score giving the percentage of SNPs called for that sample.

Cases and controls

Throughout this thesis, cases are those patients with the phenotype of interest, whereas controls are those without the phenotype.

GenCall

A sample (genotype) and SNP specific quality score that measures the distance of an individual genotype position from the centre of the cluster for that genotype and SNP. GenCall scores are calculated for each genotype, and give an estimate of how reliable that genotype is (for a specific sample and SNP). It ranges from 0 to 1, with a lower score indicating increasing distance from the centre of a cluster, implicating worse clustering.

Genetic models

All models were coded to determine the impact of the minor allele.

Genotypic - each genotype forms a separate group.

Allelic - the allelic odds ratio, derived from the counts of major and minor alleles in the case and control group.

Dominant - major allele homozygotes versus the remaining genotypes grouped together.

Recessive - minor allele homozygotes versus the remaining genotypes grouped together.

GenTrain

A SNP specific statistical score, defining how well an individual SNP has assayed, taking into account the shapes of the clusters and their relative distance to each other. The GenTrain scores can range from 0 to 1, with 0 representing poorly performing/failed SNPs.

Germline

Normally the DNA content of germ cells, in this thesis taken to mean the DNA content of non-tumour cells (cf. somatic), but also the unaltered DNA content of tumour cells.

Illumina naming convention

The alleles for each SNP are coded A and B, the A allele of any SNP is always the adenosine (A) or thymine (T), the B allele always cytosine (C) or guanine (G) . All SNPs on the BeadChip are either A or T substituted by either C or G, there are no A/T or C/G SNPs present on the BeadChips (Table 2-3).

Infinium II

The Illumina proprietary chemistry used for genotyping in the Hap300 and Hap370 arrays in this thesis.

Manhattan plot

A plot of the $-\log_{10}$(p-value) of a series of SNPs against the position on the chromosome.

Quotes

Quotes are indicated in italics.

References

References adjacent to numbers are indicated by (ref), unless when next to a genetic locus or gene, when they are indicated in numerical superscript as usual.

Survival endpoints

Disease-free survival (DFS) - interval from start of therapy to relapse or new CRC diagnosis (stage 1-3 only).

Progression-free survival (PFS) - interval from start of therapy to progression (stage 4).

Where a term refers to relapse or progression in a cohort of all stages, DFS is used to describe the time from therapy to either endpoint.

Overall survival (OS) - interval from start of therapy to death from any cause.

CRC specific overall survival - interval from start of therapy to death from CRC

Somatic

Normally the DNA content of somatic cells, in this thesis taken to mean the DNA content of tumour cells when used in the term somatic changes or markers.

Chapter 1

Introduction

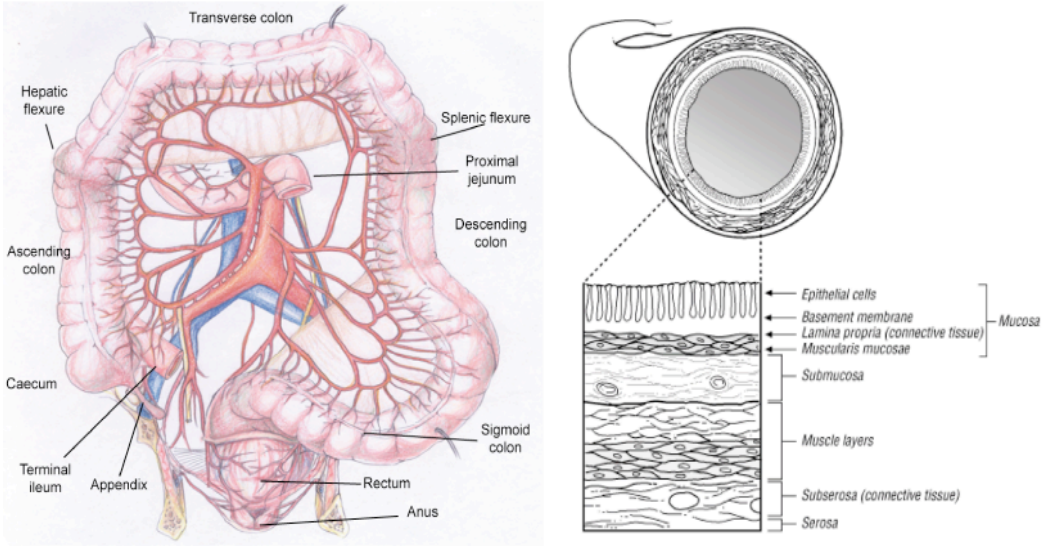
Colorectal cancer (CRC) is an epithelial malignancy of the large bowel. It represents a global health burden, with an estimated 1 million new cases per year, with over half a million deaths, resulting in an estimated mortality rate of 52%. The majority of these cancers occur in the developed world, 64% for men and 66% for women, suggesting environmental and possibly genetic factors that influence CRC incidence[2]. While treatments have improved, the disease-specific mortality, even in modern Western healthcare settings, remains at 33%, accounting for about 10% of all cancer deaths[3]. A better understanding of why some patients with CRC have different outcome from others, and why they develop CRC in the first place, will be of major benefit in dealing with this deadly disease.

1.1 Anatomy of the large bowel

The large bowel is the terminal part of the alimentary tract, extending from the ileo-caecal valve to the anal verge. It consists of several parts: caecum, ascending colon, transverse colon, descending colon, sigmoid colon, rectum, with a largely uniform structure (Figure 1-1). Apart from the distal sigmoid colon and rectum, the whole of the large bowel lies intraperitoneally. The wall of the large bowel is made up of several layers: the innermost layer is the mucosal surface with a monolayer of columnar epithelium and mucus secreting goblet cells, supported by a basement membrane. The epithelial surface invaginates, forming colonic crypts (of Lieberkühn). Next is the *lamina propria* made up of connective tissue, the mucosal neurovasculature and lymphoid tissue, followed by the *muscularis mucosae*, a thin layer of smooth-muscle cells. These layers form the *mucosa*. Next comes the *submucosa* with larger vessels and loose connective tissue, supported by the thicker of the two muscular layers, the *muscularis externa*. The surface of the large bowel 'facing' the abdominal cavity, the peritoneal surface, is covered by a monolayer of non-keratinising squamous epithelium, the serosa. Between serosa and *muscularis externa* lies a

further thin layer of loose connective tissue[4]. The overwhelming majority of CRC arise from the columnar epithelium to form adenocarcinomas, and in this thesis, colon cancer, rectal cancer and CRC are taken to mean such malignancies.

Figure 1-1 Anatomy of the large bowel



The major parts of the large bowel from the proximal colon at the ileo-ceacal valve to the anus, which as it has a different epithelial layer is not part of the large bowel (left). Cross-section of the colon depicting the anatomical layers from mucosa to serosa (right). (Images reproduced with permission: left by Kathy Mak[5] (modified); right obtained from http://oreilly.com/medical/colon/news/equivalents.html).

1.2 Demographics and survival of CRC

Colorectal cancer is a disease of the older population with an average age of onset of 66 years[3], and although the incidence has been falling, it still is between 45 and 50 per 100,000 persons per year in the Western world. It affects men slightly more than women 56.0/100,000 compared to 35.7/100,000, giving a total of 36,500 new cases per year in the UK. It contributes roughly equal proportions to the overall cancer burden of the sexes, about 10% each. The overall mortality from CRC irrespective of stage (Table 1-3) is 43.4% in the UK, compared with 33.6% the US, although the gender picture here is reversed compared to incidence, with a mortality of 42.2% for men compared to 44.8% for women in the UK, and 31.4%, and 35.9% in the US, respectively. Due to the lower incidence rate, the actual mortality rate per 100,000 women is lower than that for men: 15 compared to 23 (UK figures: CRUK Cancer Stats[6], US figures: 2008 Cancer Statistics[3]).

Over the past ten years, the improvements in CRC diagnosis and therapy have led to a steady fall in mortality, from a peak of 30 per 100,000 persons in the early 1970s to 18 per 100,000 in 2005, with the benefit driven equally by falls in male (35 to 23 per 100,000) and female (26 to 15 per 100,000) mortality. This improvement in survival has been

particularly marked in the younger age groups, but even the over 80s have experienced a 10% drop in mortality over the last 10 years[6].

1.3 Risk factors for CRC

1.3.1 Lifestyle factors

In most cases, a clear aetiological association with either a particular risk factor or genetic pre-disposition is not apparent, making CRC largely a sporadic disease. Nonetheless, in addition to the inherited predisposition (section 1.3.2), there is a clear evidence that some environmental or life-style factors increase the risk of CRC.

Immigrants from low risk areas moving to high risk areas assume the risk of the 'recipient' area within the first offspring generation: in Chinese migrants to the US, CRC risk increased with number of years lived in the US[7], while the CRC rates increase the more the diet resembles a Western diet rich in red meat, animal fat and low in fibre[8]. The protective properties of a high fibre diet, postulated because of the low CRC incidence in countries with a traditionally high fibre diet[9] have been challenged by large, prospective trials where it was not independent of the increased risk attributable to red meat intake[10,11]. Ultimately, fibre intake may merely reflect fruit and vegetable intake, which may be inversely related to red meat consumption.

Several large case-control studies have found that total calorie intake, sedentary lifestyle and being overweight also combine to increase CRC risk, in men more than women[7,12], while in women, the waist-to-hip ratio may also be important, with body fat carried predominantly around the waist conferring a higher risk[13].

Lastly, tobacco smoke has been linked to CRC risk, with increasing cigarette consumption conferring a higher risk initially of adenoma, but later also carcinoma, both in men[14] and women[15].

1.3.2 Genetic predisposition

At least one third of all CRC is thought to have a heritable component[16], but only about 5% are explained by the high-penetrance familial syndromes[17]. These syndromes, however, give an insight into some of the genes important in colorectal carcinogenesis.

1.3.2.1 HIGH-PENETRANCE SYNDROMES

1.3.2.1.1 Familial Adenomatous polyposis

Familial adenomatous polyposis (FAP) is an autosomal dominant disease caused by mutations in the Adenomatous polyposis coli gene (*APC*) located on chromosome 5. Originally described in 1934 as a hereditary disease causing multiple adenomata in the large bowel, it is not the most common of the CRC predisposition syndromes, but perhaps the best known[18]. The locus for FAP was mapped to chromosome 5q21 by

linkage analysis[19] and cloned a few years later by two groups independently[20,21]. APC is a large protein (316KDa, 2,843 amino acids) with many different mutations, although clustered around two mutational hotpots (codons 1061 and 1307). The vast majority of mutations are truncating, leading to non-functioning gene product, with a small number of missense mutations[22]. The incidence is about 1 in 8000, accounting for 0.5%-1.0% of all CRC, and making it one the most common dominantly inherited diseases[23]. Affected individuals have a large bowel carpeted with hundreds or thousands of colorectal polyps by the second or third decade, and by the age of 40, almost invariably develop frank carcinoma[24]. In addition, extra-colonic features include small bowel adenomata and (occasionally) carcinoma, Congenital Hypertrophy of the Retinal Pigment Epithelium (CHRPE), and desmoids tumours[25]. A recessive form of colonic polyposis is associated with germline mutations in *MYH*[26].

1.3.2.1.2 Hereditary Non-polyposis Colon Cancer (HNPCC)

Hereditary Non-polyposis Colon Cancer is an autosomal dominant disorder due to mutations in the mismatch repair (MMR) genes *MLH1*, *MSH2*, and less commonly in *MSH3*, *MSH6*, and *PMS2*[27,28]. These genes function in the mismatch repair of spontaneous errors during DNA replication in DNA (as well as in response to exogenous DNA toxins), most commonly seen in short mono- or dinucleotide repeats (microsatellites), thus maintaining the fidelity of DNA replication[29]. HNPCC should be suspected if a set of clinical criteria are met, e.g. the modified Amsterdam criteria[30], and diagnosed by subsequent demonstration of a mutation in an MMR gene. The penetrance of HNPCC is about 80%, and extra-colonic cancers occur in a subsets (particularly endometrial cancer), and in the case of skin tumours, is termed Muir-Torre syndrome. HNPCC accounts for about 1-3% of all CRC, depending on the population studied[31,32].

Table 1-1 Diagnosis of HNPCC

Modified Amsterdam criteria for genetic testing for HNPCC
At least three relatives with HNPCC-associated cancer (colorectal, endometrial, small bowel, ureter or renal pelvis)
One should be a first-degree relative of the other 2
At least two successive generations should be affected
At least one case before age 50
FAP should be excluded
Tumours should be verified histopathologically

Adapted from Vasen *et al.*[30]

1.3.2.1.3 Hamartomatous Polyposis Syndromes

Mutations in *LKB1*, a serine threonine kinase on chromosome 19, cause Peutz-Jeghers Syndrome (PJS), an autosomal dominant disorder[33]. LKB1 has an essential role in the G1 cell cycle arrest. The classical clinical lesion in PJS are pigmented macules of the lips and mucous membranes of patients who also are prone to multiple gastrointestinal

hamartomatous polyps with an increased risk of gastrointestinal as well as other cancers. Juvenile Polyposis (JP), also autosomal dominant, has a similar clinical presentation to PJS, although polyps are more commonly colonic. The classical causal mutation is in *PTEN*[34], although mutations in *SMAD4* and *BMPR1* have also been described[35,36]. The polyps in the hamartomatous polyposis syndromes have malignant potential, although they account only for a minority of CRC.

1.3.2.2 COMMON VARIANTS

If, as stated above, only a proportion of the inherited predisposition to CRC can be explained by the high-penetrance syndromes described in section 1.3.2.1, other loci must play a part. It is unlikely that further high penetrance syndromes exist that contribute appreciably to the overall familial CRC burden, even if this does not exclude the existence of further rare high penetrance loci. The common variant - common disease hypothesis suggests that there are many common variants that individually increase the risk of the individual only by a small fraction, in other words have a low penetrance. Due to their common nature, minor allele frequencies (MAF) of 5-49%, they have the potential to contribute substantially to familial CRC. With increasing knowledge of the structural variants in the human genome, in particular the abundance of single nucleotide polymorphisms (SNPs), and the ability to perform highly multiplexed assays to interrogate these loci, genome-wide association studies (GWAS) have become feasible, and have led to a marked increase in the understanding of the genetic architecture of many complex disease traits[37]. GWAS using high-density SNP arrays have been successful in the identification of susceptibility loci for colorectal and other cancers[38,39], as well as non-malignant disease[40,41].

In CRC, 10 loci (13 SNPs) have been described (see Table 1-2), derived from two GWAS with more than 13,000 cases and controls and over 27,000 verification subjects[42-46]. None of these SNPs had previously been linked to CRC development, and only three SNPs in *SMAD7* on chromosome 18q lie within a known CRC predisposition gene[45], although the locus on chromosome 15 with two SNPs had been linked to increased CRC risk in Ashkenazi Jews[43].

Only one of the CRC susceptibility SNPs has been linked to cancer susceptibility in a different tissue type: rs6983267 on chromosome 8q24 is also associated with an increased risk of prostate cancer[38], and while variants on 8q24 also confer a higher risk of breast, ovarian and bladder cancer, rs6983267 was shown not to be the associated variant in those instances[47-49]. The initial hypothesis was that the transcription factor *POU5F1P1* was the gene tagged by rs6983267, but the proto-oncogene *MYC*, implicated in pathways important to the development of colon[50], breast[51] and other cancers lies 337Kb telomeric

and could harbour the true causal variant(s) detected by long-range LD. It is possible that the region harbours several tissue specific enhancers of *MYC*, defined by the five LD blocks in this region: one associated with CRC and prostate cancer, containing rs6983267, one associated with breast cancer, and three associated with prostate cancer alone[49]. Thus, as expected from studies that directly type only a tiny fraction of the sequence variation in the human genome, the causal variants within 8q24 have yet to be elucidated.

The ten regions identified so far, based on the assumption of an additive model, collectively probably account for approximately 6% of the excess familial risk in CRC[46].

Table 1-2 Known CRC low-penetrance risk loci

SNP	Alleles	Region	OR (95% CI)	p-value	Gene	Distance
rs6983267[44]	T/G	8q24	1.21 (1.15-1.27)	1.3E-14	*MYC*	337Kb
rs16892766[42]	A/C	8q23	1.25 (1.19-1.32)	3.3E-18	*EIF3H*	27Kb
rs10795668[42]	A/G	10p14	0.89 (0.86–0.91)	2.5E-13	none	within 500Kb
rs3802842[52]	A/C	11q23	1.17 (1.12-1.22)	1.08E-12	*POU2AF1*	51Kb
rs4444235[46]	T/C	14q22	1.11 (1.08–1.15)	8.1E-10	*BMP4*	9.4Kb
rs4779584[43]	T/C	15q13	1.26 (1.19-1.34)	4.4E-14	*GREM1*	15Kb
rs9929218[46]	A/G	16q22	0.90 (0.87–0.94)	1.2e-08	*CDH1*	intronic
rs4939827[45]	T/C	18q21	0.85 (0.81–0.89)	1.0E-12	*SMAD7*	intronic
rs10411210[46]	T/C	19q13	0.83 (0.78–0.88)	4.6e-09	*RHPN2*	intronic
rs961253[46]	A/C	20p12	1.12 (1.08–1.16)	2.0E-10	*BMP2*	342Kb

MYC – *MYC* proto oncogene; *EIF3H* - eukaryotic translation initiation factor 3; *POU2AF1* - POU class 2 associating factor 1; *BMP* - bone morphogenic protein; *Grem1* - gremlin-1; *CDH1* - E-cadherin; *RHPN2* - Rho GTPase binding protein 2. Where more than one SNP has been described per locus, only the most significant SNP is listed.

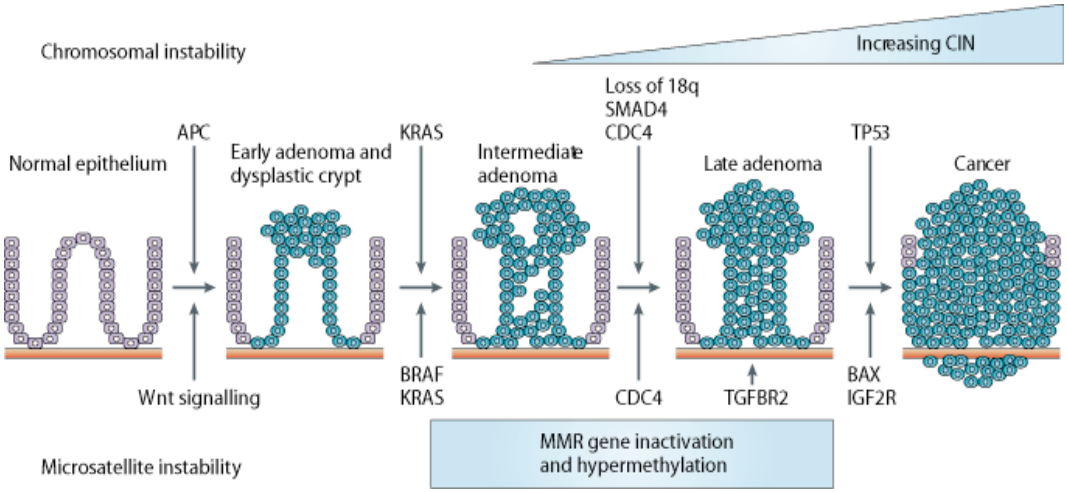
1.4 Colorectal tumorigenesis

Colorectal cancers are characterised by genomic instability, either chromosomal instability (CIN) with many cytogenetic abnormalities, occurring in the majority of CRC, or microsatellite instability (MSI), with mutations in simple nucleotide repeats (microsatellites) throughout the genome[53]. CIN is present in about 65-70% of CRC, and in practice, often inferred from finding aneuploidy and/or polyploidy on flow cytometry[54]. About 15% of CRC have a near-diploid chromosome set and MSI, defined as a tumour having instability in at least two of five standard microsatellite markers[55]. While about one third of CRC do not display either form of instability[56], direct measurement of aneuploidy utilising flow cytometry is relatively crude and it is likely that the 'double-negative' group harbours some cytogenetic abnormalities not detected by cytometry[57].

CRC also harbour epigenomic instability, either as global hypomethylation or as the CpG Island Methylator Phenotype (CIMP), defined as methylation at 3 or more specific

marker loci[58]. The latter has considerable overlap with MSI, and CIMP appears to be associated with MSI in patients who do not harbour germline mutations in mismatch repair (MMR) genes, which are characteristic of Hereditary Non-Polyposis Colon Cancer (HNPCC or Lynch syndrome)[58]. In addition, this group is strongly associated with *BRAF* V600E mutations (previously named V599E[59]), and has been variably termed CIMP-high[60,61] (CIMP-H) or CIMP 1 (ref [62]). A third group of the CIMP+ve phenotype is associated with *KRAS* mutations and has been termed CIMP-low[63] (CIMP-L) in response to the not universally reported[60] lower levels of methylation, or CIMP 2 (ref [62]); see also section 1.6.1.5 and Figure 1-6.

Figure 1-2 Adenoma-carcinoma sequence



The adenoma-carcinoma sequence, first described by Vogelstein *et al.*[64], see text for more information. Picture modified with permission from Walther *et al.*[1]

Tumorigenesis leading to CRC with chromosomal instability is said to occur in a series of steps involving specific mutations in key genes along the way, the Adenoma-Carcinoma model initially proposed by Vogelstein *et al.* in 1988 (ref [64]). This model is likely to be an oversimplification, but aligns observed clinico-pathological changes with genetic abnormalities in the progression of chromosomally unstable CRC (the "gatekeeper pathway" involving genes that regulate cell growth[65]). The initial step in tumorigenesis is that of adenoma formation, associated with loss or somatic mutation of *APC*. Larger adenomas and early carcinomas acquire mutations in the small GTPase, *KRAS*, followed by loss of chromosome 18q with *SMAD4*, downstream of TGFβ, and mutations in *TP53* in frank carcinoma (Figure 1-2).

1.4.1 Genetic changes leading to CRC with chromosomal instability

1.4.1.1 *APC*

Mutations of *APC* are present in at least 60 – 70% of CRC[66,67], and *APC* appears to be involved in at least two independent pathways important to colorectal tumorigenesis.

Firstly, APC acts in the degradation of β-catenin, a molecule involved in cell-cell adhesion by anchoring E-cadherin to the intracellular cytoskeleton. Free β-catenin is phosphorylated by a complex consisting of APC, axin and glycogen synthase kinase-3β (GSK), which marks it for ubiquitination and subsequent degradation in the proteasome. In the presence of extracellular WNT bound to its receptor, Frizzled, the latter phosphorylates the cytosolic protein Dishevelled which in turn leads to the dissociation of the APC/axin/GSK complex and the accumulation and translocation to the nucleus of free β-catenin. Here β-catenin interacts with T-cell factor (TCF) leading to the transcription of genes involved in cell cycling, including cyclin D1. More than 90% of patients with CRC have alterations affecting this pathway[68].

More recently, a role for APC in the mitotic spindle checkpoint has been suggested: *APC* mutations cause defects in microtubule plus-end attachments leading to CIN in cell lines[69]. Introduction of mutant APC into diploid cell lines leads to CIN[70], and resistance to pharmacological spindle disruption[71]. Mouse embryonic stem cells with homozygous *APC* mutant genotype show extensive chromosome and spindle aberrations[72,73]. Lastly, *APC* mutations are less common in diploid, MSI+ tumours[74] than in CIN+ tumours[75,76], although some diploid cell-lines and CRC have mutant APC (Figure 1-3).

1.4.1.2 *KRAS*

The *KRAS* (Kirsten-ras) gene is located on chromosome 12 and its product plays a central role in the signal transduction of several pathways (e.g. epidermal growth factor (EGF)[77], vascular endothelial growth factor (VEGF)[78], platelet-derived growth factor (PDGF)[79], fibroblast growth factor (FGF)[80]) by converging these onto the mitogen-activated protein kinase (MAPK) signalling pathway. Following the binding of the extra-cellular signalling molecule to its respective receptor, the associated receptor-tyrosine kinase becomes activated and recruits adaptor proteins (Grb2 and SOS) to activate KRAS. In a phosphorylation cascade, this activates RAF, then MAPK-kinase (MEK), and then the MAP kinases ERK1 and ERK2 in successive steps. The ultimate outcome of this cascade is an anti-apoptotic signal[81] and, via cyclin D1, regulation of the cell cycle[82], although the precise effect on the cell cycle depends on the duration of the signal[83]. Consequently, activating mutations of *KRAS* generally drive cellular proliferation and they are found in between 30% and 50% of CRC[84,85].

1.4.1.3 *TP53*

Germline mutations in *TP53* are the underlying abnormality of the Li-Fraumeni syndrome, a hereditary cancer susceptibility syndrome predisposing individuals to many cancers, in particular sarcoma[86]. Somatic mutations are also found in about half of sporadic solid tumours and up to 60% of human CRC[87]. *TP53* codes for a transcription

factor that in resting cells is switched off through the interaction with MDM2, mediating ubiquitin-dependant degradation by the proteasome. The best known function of TP53 is its involvement in DNA damage checkpoints during the cell cycle, with the ability to arrest mitotic division to allow repair or induce apoptosis, acting predominantly during the G_1 and G_2 DNA checkpoints[88]. It appears, however, that it can affect a much wider range of cellular functions: invasion and motility, angiogenesis, cell survival and regulation of oxidative stress, repair of genotoxic damage, and autophagy[89].

1.4.1.4 LOSS OF CHROMOSOME 18Q

The long arm of chromosome 18 is fully or partially lost in 50–70% of CRC[90], and appears to drive tumorigenesis as re-introduction of lost 18q reverses the malignant phenotype in cell-lines[91]. The likely gene is *SMAD4* (also known as *DPC4*): Mutations in *SMAD4* occur in only 6 – 35% of CRC[92-94], but loss or reduction of SMAD4 function through loss of 18q occurs in 60 – 70% of CRC[95,96]. While many CRC exhibit bi-allelic loss[97], even loss of heterozygosity (LOH) may play a role in colonic tumorigenesis[98].

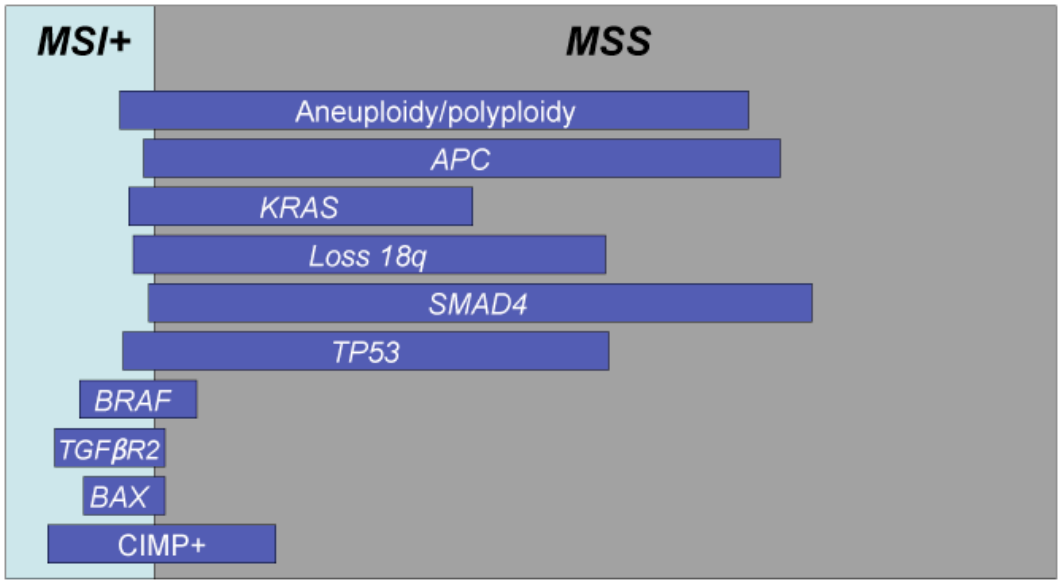
SMAD4 functions in the signal transduction of the transforming growth factor β (TGFβ) pathway: after binding of TGFβ to the heterodimeric TGFβ-receptor (TGFβR1 and TGFβR2), SMAD2 (also located on 18q21) and SMAD3, so-called receptor-activated or R-SMADs, are phosphorylated and form a complex with SMAD4 (a common mediator or Co-SMAD), before translocating to the nucleus and interacting with a variety of transcription factors and modulators[99,100]. The interaction with other proteins at the *p15^{INK4B}* promoter leads to upregulation of p15^{INK4B} and subsequent cell cycle arrest[101]. Together, these mechanisms form a potent tumour suppressor pathway[97,102]. TGFβ signalling can be repressed by the "inhibitory" SMADs 6 and 7, which are part of a TGFβ induced negative feedback loop. Thus, loss of function mutations in either *TGFβR2* or *SMAD4* will decrease the inhibitory signal and can lead to tumorigenesis. However, signalling through the TGFβ pathway is more complicated than this linear model suggests, and 29 members of the TGFβ ligand superfamily as well as five isoforms of receptors 1 and 2 have been described. The receptors can form homo- and heterodimers which themselves can combine with a variety of R-SMADs and Co-SMADs to mediate the different effects of the various ligands[103]. Further, some of the downstream outcomes actually facilitate tumour progression and invasion through a transition from epithelial to mesenchymal morphology of the tumour cells[104,105]. Lastly, TGFβR signalling can effect changes in SMAD independent pathways, such as MAPK, Erk and JNK, which may be growth promoting.

1.4.2 Genetic changes leading to CRC with microsatellite instability

Microsatellite instability is found in about 15-17% of CRC, of which only a minority are HNPCC tumours. MSI+ve CRC, characterised by deficiency of the mismatch repair system, leading to 'slippage' in microsatellites (the "caretaker pathway" involving genes maintaining genomic stability[65]), only carry the changes described for CIN+ve CRC in section 1.4.1 infrequently, thus development of CRC must involve different but analogous genetic changes (Figure 1-2).

The predominant impairment of MMR genes in sporadic MSI+ve CRC is the down-regulation of MLH1 expression through promotor methylation[106], although the MSI status is increased through positive selection of tumour cells with mutated microsatellites in *MSH3* and *MSH6*[107].

Figure 1-3 Frequency of mutations seen in MSI+ve and MSI-ve CRC



Data for each marker are derived from the several sources (where available). All studies have at least 200 patients unless only one study citing incidence of genetic change across MSI spectrum. Aneuploidy[108]; *APC*[68,75,109]; *KRAS*[61,87,110,111]; loss of 18q[56,112,113]; *TP53*[87,111,114]; *SMAD4*[95]; *BRAF*[60,61,115,116]; *TGFβR2*[56,117]; *BAX*[118]; *CIMP*[60,61,115,119,120]

However, MSI is uncommon in adenomata[121], and the initial step is thought to involve alteration in *WNT* signalling[28], possibly involving *AXIN*[122]. *BRAF* mutations, common in MSI+ve CRC, are likely to occur in the place of *KRAS* mutations[123], although the latter do occur in a minority. Further positive selection occurs for mutations affecting microsatellites in transforming growth factor receptor 2 (*TGFβR2*)[124], caspase-5 (ref 125), and insulin like growth factor 2 receptor (*IGF2R*)[126]. These genes appear to have mutations in both alleles without cancer cells displaying a similar rate of mutation in other microsatellite tracts of similar length[107,124]. Mutations in *TGFβR2* induced by MSI lead to the disappearance of the TGFβR2 from the cell surface and an escape from TGF-beta-

mediated growth control[127], with analogous consequences to loss of SMAD4 function in CIN+ve CRC. While *TP53* mutations are rare in MSI+ve CRC, mutations in the pro-apoptotic gene *BAX* are common, providing a *TP53* independent mechanism for the transition from late adenoma to carcinoma[128].

CDC4 inactivation may precede *TP53* mutation[129], leading to increasing CIN[130], although it is not always associated with CIN and may also play a role in the MSI pathway[131].

1.5 Treatment of CRC

Treatment recommendations for CRC are based on pathological stage. While several staging systems are in clinical use, the oldest being Dukes' system described in 1932 for rectal cancer[132] and subsequently modified to include the colon and distant metastatic disease[133], the recommended staging system is the TNM system[134-136].

The TNM system is shown in Table 1-3 assesses the tumour itself (T) for its depth of invasion into the bowel wall, the lymphnode status (N) according to the number of lymphnodes involved, and the presence of distant metastasis (M). CRC staged by the TNM system are then often grouped into categories with similar prognosis, the AJCC stages (Figure 1-4 and Table 1-4).

Table 1-3 Description of the components of the TNM system

Stage	Code	Description
Tumour	Tx	Primary tumour cannot be assessed
	T0	No evidence of primary tumour
	Tis	Carcinoma in situ: intraepithelial or invasion of the lamina propria
	T1	Tumour invades submucosa
	T2	Tumour invades muscularis propria
	T3	Tumour invades through the muscularis propria into the subserosa or into non-peritonealised pericolic or perirectal tissues
	T4	Tumour directly invades other organs or structures and/or perforates visceral peritoneum
Lymphnode	Nx	Regional nodes cannot be assessed
	N0	No regional lymph node metastasis
	N1	Metastasis in 1 to 3 regional lymph nodes
	N2	Metastasis in 4 or more regional lymph nodes
Metastasis	Mx	Distant metastasis cannot be assessed
	M0	No distant metastasis
	M1	Distant metastasis

The TNM staging system assesses the depth of tumour invasion, the number of involved lymphnodes, and the presence of distant metastasis [134].

For stage 1 CRC, the only treatment required is surgical resection, and appropriate follow-up with regular colonoscopy. There is no evidence for any benefit from adjuvant therapy and 5-year survival rates are in excess of 93% for correctly classified stage 1 CRC[135]. The picture is similar for stage 2 CRC, for which the role of adjuvant systemic therapy after curative resection remains controversial, and the added survival benefit at 5 years is 3-4 %[137] with 5-fluorourcail (5-FU) alone, and a further 2% with the addition of oxaliplatin[138].

For stage 3 CRC, the role of systemic adjuvant combination chemotherapy is well established: the relative reduction in risk of death from CRC with 5-FU alone estimated at 26%[139], and the addition of oxaliplatin reduces the risk of death from CRC at 5 years by a further 10%[138]. Despite good activity of irinotecan in the metastatic setting, there is no role for it in the adjuvant setting after three trials failed to demonstrate survival benefit[140-142]. There is currently no role either for the use of targeted agents in the management of AJCC stage 3 CRC following the negative results of the NSABP C-08 study[143], with further trials still underway testing this hypothesis[144,145].

Within the early stage groups, in particular stage 2 and 3, where adjuvant therapy must be considered, histopathological examination of tumour material can help to define prognosis further[146], using lympho-vascular invasion[147]; resection margins (as part of the TNM staging); and tumour grade[148]. In addition, some clinical parameters independently appear to influence outcome: obstruction/perforation at presentation[149]; performance status[150]; and pre-operative carcinoembryonic antigen (CEA) levels[151], which probably reflect tumour burden at the time of surgery. Patients with stage 2 CRC and adverse histological factors are deemed high-risk and treated along the same lines as stage 3 patients.

Table 1-4 Stage groupings based on TNM system

TNM	AJCC 5th ed	AJCC 6th ed	Dukes	Astler-Coller
T1 or T2, N0, M0	I	I	A	A, B1
T3, N0, M0	II	IIA	B	B2
T4, N0, M0	II	IIB	B	B3
T1-2, N1, M0	III	IIIA	C	C1
T3-4, N1, M0	III	IIIB	C	C2, C3
Any T, N2, M0	III	IIIC	C	C1, C2, C3
Any T, Any N, M1	IV	IV		D

Stage groupings according to the AJCC. The latest edition (6th) subclassifies stage 2 and 3 CRC to take account of differing prognoses in these subgroups [134].

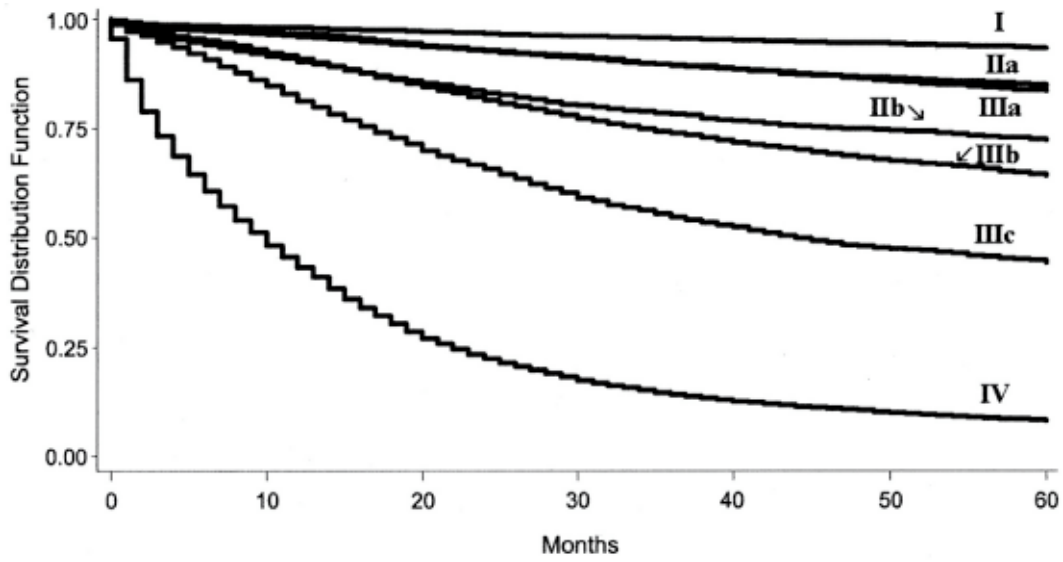
In stage 4 CRC, 5-FU in combination with either oxaliplatin or irinotecan is the mainstay of therapy, with potentially curative surgery considered for those who have limited sites of metastatic disease and a good response to chemotherapy[152]. Increasingly, targeted agents

are finding their way into the management of stage 4 disease, with in particular monoclonal antibodies directed against the EGF receptor (EGFR) and VEGF being utilised[153,154].

1.6 Genetic markers of outcome in CRC

Despite the overall improvements in CRC therapy, our understanding of why individual patients do or do not respond to therapy or relapse remains poor. Consequently, there is a large, unknown subset of patients who receive treatment from which they do not benefit. In the metastatic setting, this failure to benefit is relatively easy to judge by demonstrating tumour progression using clinical and radiological assessment. In the adjuvant setting, many patients must be treated with significant attendant toxicity[155] so that relatively few will benefit as there are clearly patients who would not have relapsed even without adjuvant therapy[156]. Understanding the reasons for treatment failure, and developing an ability to predict those who would benefit the most (and least) remain important aims in the management of CRC.

Figure 1-4 CRC survival by stage



Kaplan Meier curves for survival in CRC by stage (from O'Connell *et al.*[135]). It is clear from this graph that on a population basis, clinico-pathological staging does sub-divide patients into groups with similar prognosis. It is also clear that the current staging systems are imperfect, with stage IIB having a worse prognosis than stage IIIA.

To date, the gold standard for prognostication remains clinico-pathological staging as described in section 1.5. While clinico-pathological staging separates patients into groups with distinct outcomes, importantly, it offers little information about response to treatment in individual patients, or about the risk of relapse in early stage CRC.

A number of protein and genetic markers have been described in an attempt to refine prognostic information and predict benefit from systemic treatment in an attempt to spare

those who will not benefit from the toxicities associated with systemic therapy. None of these are in routine clinical use[157], but much excitement has been generated by the progress made in other tumour types towards such goals. In breast cancer, for example, hormone receptor status is associated with prognosis[158] and response to hormonal therapies[159], there is a clearer understanding of who benefits from trastuzumab therapy[160], and a gene expression signature conveying worse prognosis[161] has been approved by the FDA to support clinical decision making[162]. While the CRC community has lagged behind somewhat, there is now a growing knowledge base of analogous determinants.

1.6.1 Hypothesis driven markers

To date, the most studied markers are somatic (acquired) changes for which some biological rationale exists for a potential impact on cancer outcome. These are often changes associated with tumour progression in the adenoma-carcinoma sequence model (Figure 1-2) or the observed type of genomic instability (chromosomal or microsatellite). Where germline (inherited) changes have been studied, they have been in pharmacological pathways involved in the metabolism and mechanism of action of 5-FU, the therapeutic mainstay for CRC. They are all hypothesis driven, with often limited *a priori* evidence to suggest a link to prognosis.

1.6.1.1 *KRAS*

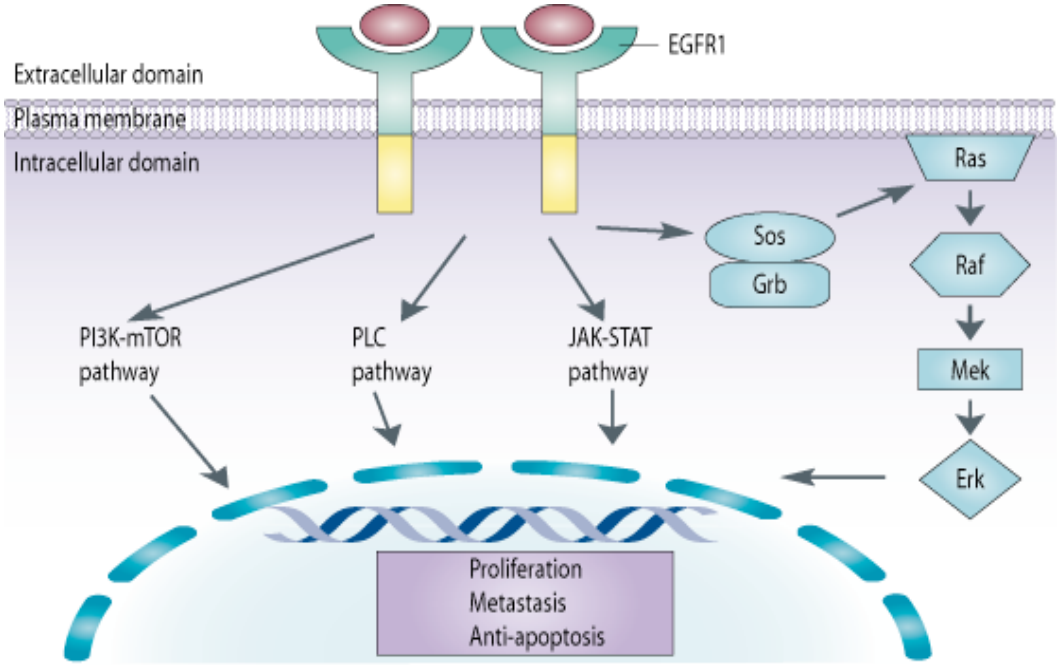
Many studies have evaluated *KRAS* mutations in exon 2 (codons 12 and 13) and to a lesser extent in exon 3 (codon 61), for their association with outcome in CRC in general. *KRAS* mutations are an early event in the adenoma-carcinoma sequence, although they are demonstrable only in about a third of CRC, with a large majority located in codon 12[163]. Mutations in all three codons compromise the ability of GTPase activating proteins to effect the inactivating hydrolysis of KRAS bound GTP to GDP[164]. Other mutations are uncommon as they result in lower constitutive KRAS signalling than mutations in codons 12, 13, and 61 and are selected against in the tumour[165].

While the largest international effort to combine data from different groups, the RASCAL collaborative, found that mutations generally confer a worse prognosis[163], the subsequent analysis including a further 700 patients found only the glycine to valine substitution at codon 12 to be associated, and only in AJCC stage 3 patients[84]. Other studies have found no association, and to date, there is no convincing evidence to suggest that *KRAS* mutations are independent prognostic factors in CRC.

More recently, however, KRAS mutation status has been established as a predictive marker for treatment with EGFR inhibitors. This interaction was initially observed with small molecule inhibitors of EGFR in non-small cell lung cancer (NSCLC)[166], but several studies in CRC have now shown that due to the convergence of the EGFR and KRAS

pathways (Figure 1-5), patients with stage 4 *KRAS* mutant CRC receiving treatment with anti-EGFR antibodies, cetuximab[167-171] and panitumumab[172] derive significantly less benefit than patients whose tumours are *KRAS* wildtype. Further, in 80 *KRAS* wildtype patients treated with cetuximab, 11 patients with the *BRAF* V600E mutation did not respond[173]. BRAF acts downstream from KRAS and mutations in the two appear mutually exclusive[123]; *BRAF* could therefore be a locus for a second hit affecting the same pathway, and both are important determinants of resistance to anti-EGFR therapies.

Figure 1-5 KRAS and its interactions with other signalling pathways



Upon ligand binding, the EGFR type 1 homo-dimerises leading to the activation of the intracellular kinase domain. Via small adaptor proteins, the KRAS signalling cascade is activated leading to increased proliferation. Downstream of KRAS is BRAF, explaining why non-constitutively activated KRAS and BRAF are necessary for EGFR blockade to work. The initial assumption that the common EGFR overexpression in CRC[174] is also required, making this model very elegant, is challenged by a lack of correlation between EGFR expression and anti-EGFR antibody response[153] and the observation that patients with irinotecan refractory, EGFR negative CRC can still respond to irinotecan and cetuximab[175]. Picture modified with permission from Walther *et al.*[1]

KRAS is close to the ideal predictive biomarker: most mutations are limited to a small part of the gene and relatively easily detected, the negative predictive value is high (99% of patients with mutated *KRAS* do not respond to EGFR inhibition[167], even if the positive predictive value remains poor), and the effects are based on a plausible biological rationale. The example of *KRAS* further demonstrates how our evolving knowledge of cancer biology can refine treatment strategies and interventions, but also highlights the difficulties that the regulatory agencies will face when evaluating anti-EGFR or other targeted agents in response to retrospective clinical studies.

1.6.1.2 APC *AND* β-CATENIN

The *APC* gene acts to promote the degradation of β-catenin and so limits the transcription of *WNT* target genes involved in regulating the cell cycle. This pathway is integral to colorectal tumorigenesis and more than 90% of patients have alterations affecting it[68]. Given the frequency of changes, it is not surprising that neither APC nor β-catenin is a prognostic marker able to differentiate between patients. While the type of *APC* mutation could hold prognostic information – for example mutations that abolish β-catenin binding sites of the APC protein may be associated with poorer prognosis[75] – addressing specific mutations in clinical practice could be technically difficult due to the large number of *APC* mutations already described. Testing for general over-expression of β-catenin does not appear useful, but determining the cellular location of over-expressed β-catenin may hold prognostic information[176,177], even if it is likely to be a surrogate marker for a genetic change elsewhere in its degradation pathway. Changes in APC and β-catenin are thus insufficiently validated in patients and have no role in clinical practice at present.

1.6.1.3 TP53

Loss of heterozygosity at chromosome 17p is a frequent event in CIN+ve CRC and many studies have focused on the 17p region containing the *TP53* tumour suppressor gene. *TP53* has been frequently investigated as both a prognostic factor and a predictor of response to therapy with conflicting results. The methods used to assess the mutation of *TP53* vary greatly between studies, as do study designs in terms of clinico-pathological data reported and patient selection, making firm conclusions about the prognostic value of *TP53* difficult[178,179].

1.6.1.4 LOSS OF 18Q

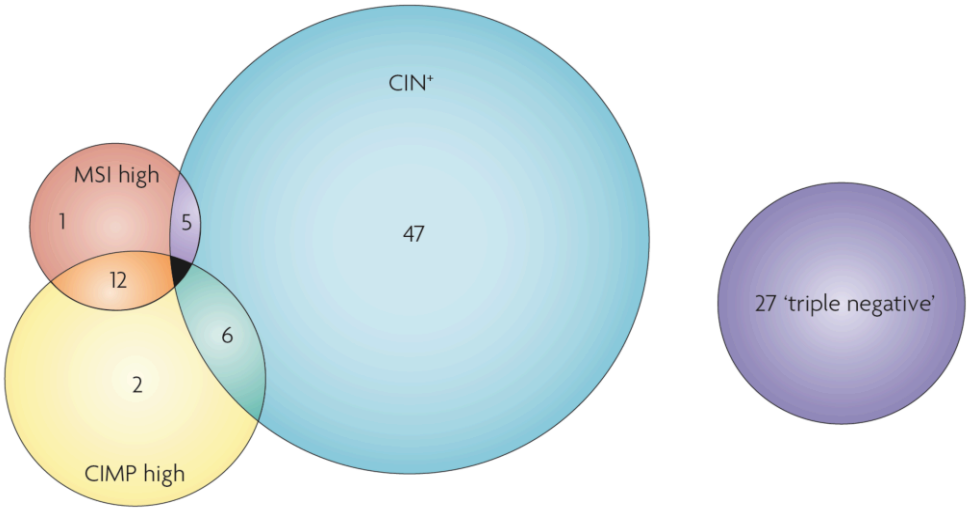
Deletion of the long arm of chromosome 18 is the most common cytogenetic abnormality in CRC and has been associated with a poorer prognosis[180], but again, this is not a uniform finding[181,182]. Many studies have investigated the genes found in this region as prognostic markers, in particular Deleted in Colon Cancer (*DCC*) without demonstrating a clear link to prognosis[183], and the known CRC predisposition gene, *SMAD4*. The latter is a member of the TGF-β signalling pathway, and decreased *SMAD4* mRNA levels appear to be associated with a worse prognosis[184] and poorer response to 5-FU[185], but not all studies find a relationship between loss of 18q and *SMAD4* expression[92], thus failing to definitively link any gene on 18q to prognosis. Further, it appears likely that loss of 18q is a marker of CIN[57] and therefore not an independent prognostic marker (section 1.6.5.1), while the other common alteration in TGF-β pathway in patients with CRC – mutations in *TGFβR2* – are almost exclusively linked to MSI[186].

Overall, to date, somatic markers have proved to be of limited prognostic utility in patients with CRC. Some of this might reflect the fact that many changes are associated with molecular phenotype, most (mutations in *KRAS*, *APC*, and *TP53*, and loss of 18q) are much less common in patients with MSI+ve CRC[113,186]. Thus, any somatic marker studied must be evaluated in the light of possible confounding by genomic instability.

1.6.1.5 (Epi) Genomic instability and prognosis

The prognostic value of CIN and MSI is no longer in question: both have been subject to large meta-analyses[187,188], which unequivocally established that patients with CIN+ve disease have a poorer prognosis (Hazard ratio for death (HR)=1.45, 95% CI 1.35-1.55, p<0.001) and patients with MSI+ve CRC have a better prognosis (HR=0.65, 95% CI 0.59-0.71, p<0.001) than patients with CIN-ve and MSI-ve CRC, respectively. Less clear is the prognostic relationship between CIN and MSI. The traditional view that they are mutually exclusive and all CRC are either one or the other is probably not correct (see Figure 1-6). While CIN and MSI may carry separate prognostic information, the only published study to date to assess the impact of both MSI and CIN in multivariate analysis did not find that the impact on survival of MSI was independent of that of CIN[108].

Figure 1-6 Overlap of CIN, MSI and CIMP



Venn diagram of CIN measured by cytometry, MSI and CIMP (all figures are percentages of the overall number of patients). About 17% of patients display MSI[188], about 60% CIN[187], and about 20% CIMP[58,60,61,115]. About a quarter of MSI+ve CRC are CIN+ve[108]. Only one study has addressed the intersection of all three forms of instability[189]. The best estimate of the distribution of the CIMP phenotype relative to CIN and MSI is that it accounts for the majority of MSI+ve, CIN-ve CRC (the sporadic MSI+ve CRC)[58], about 12% of all CRC; that CIN is independent of CIMP, and CIMP is therefore relatively evenly spread over the CIN+ve and CIN-ve groups; and that the CIN+ve MSI+ve CRC may also be CIMP+ve, but there is no evidence in the literature for this, likely to reflect the small size of this group[189]. Picture reproduced with permission from Walther *et al.*[1]

The CIMP-H phenotype has also been associated with a better prognosis compared to CIMP-ve[61], while in contrast, in MSI-ve CRC which are CIMP+ve and harbour *BRAF* V600E mutations, this phenotype is associated with a worse prognosis[190] and increasing

levels of CIMP (-ve to low to high) confer worsening prognosis in MSI-ve CRC[60]. Despite the finding in one study that MSI associated prognostic information was not independent of CIMP status[61], the relative contributions of CIN, MSI and CIMP to outcome need to be further investigated to understand the impact and interaction of these variables.

Decreased methylation in the form of global genomic hypomethylation is also involved in colorectal carcinogenesis[191] and associated with CIN[192,193], representing the opposite end of the methylation spectrum to CIMP+ve MSI+ve CRC. It has been found to be associated with worse prognosis, but this analysis did not include CIN status[192], and again, the interplay of the various types of (epi)genomic instability is not clear.

1.6.1.6 MSI, CIN AND RESPONSE TO THERAPY

Since the initial publication of a detrimental effect of adjuvant 5-FU in loco-regional MSI+ve CRC[194], further data have been published by the same group confirming their earlier findings[195]. The clinical data are supported by *in vitro* evidence showing that a functioning MMR system is required for cytotoxicity in response to 5-FU incorporation into DNA[196], in addition to its effects mediated by inhibition of thymidylate synthetase (Figure 1-7). In contrast, other retrospective data suggest that patients with MSI+ve CRC do benefit from adjuvant 5-FU[197]. This is supported by the observation that in stage 4 CRC, 5-FU is effective in MSI+ve CRC[198]. The conflicting views demonstrate the difficulty in looking at uncommon subgroups, but the evidence for a detrimental effect is sufficiently strong for a clinical trial (E5202), now underway, to compare MSI+ve CRC without adjuvant chemotherapy to MSI-ve CRC with adjuvant chemotherapy prospectively to resolve these conflicting data.

In addition, there are data to suggest that MSI+ve CRC are more sensitive to irinotecan based regimes in some clinical series[199,200] and cell line studies[201]. The increased survival following irinotecan was also observed in the adjuvant setting[199], where irinotecan had failed to be of benefit in the unselected patients[140], making it an attractive proposition for prospective study in this setting.

CIN could also be a negative predictive marker for the response to taxanes, a class of drug that showed limited activity in CRC during initial drug development. An intact spindle assembly check-point is required for taxane sensitivity, in other words, only diploid cells segregate chromosomes normally and are sensitive to paclitaxel[202,203], a hypothesis currently being tested in a phase 2 clinical trial (CINATRA) using a novel taxane, patupilone.

1.6.2 Pharmacogenetic markers

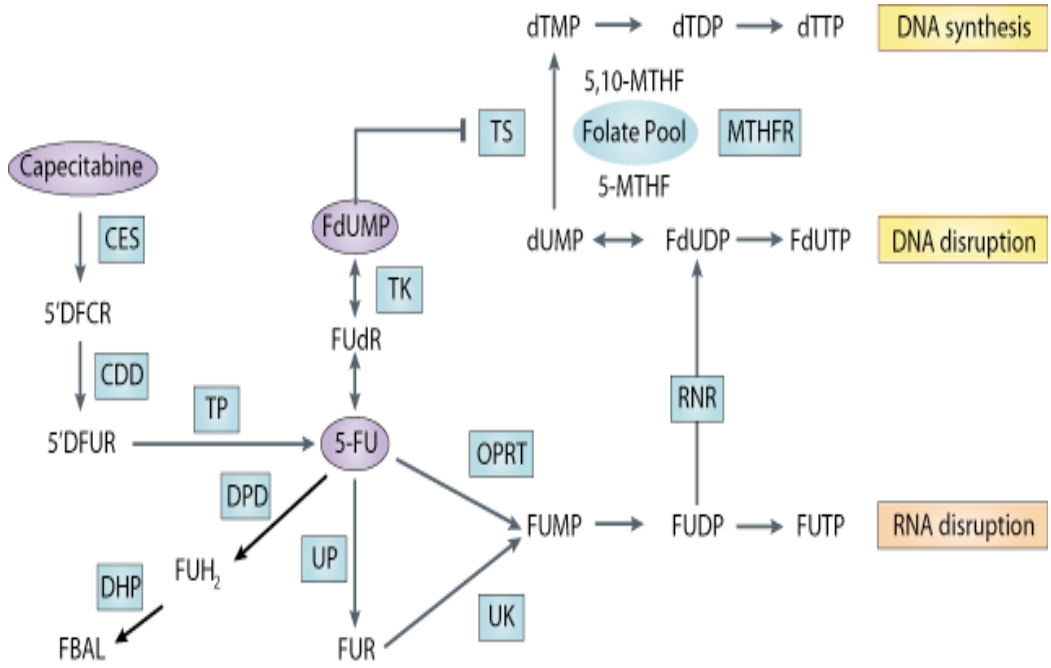
Inherited polymorphisms have the potential to impact greatly on treatment; for conventional chemotherapy agents with a narrow therapeutic window, subtle genomic

changes can modulate drug-specific pharmacokinetics or pharmacodynamics and substantially affect individual response and toxicity after chemotherapy. Almost all chemotherapy regimes for CRC incorporate 5-FU or its oral prodrug, capecitabine, often in combination with oxaliplatin or irinotecan, and the ability to tailor therapy to take account of variation in their metabolism has the potential to translate into benefit for a large patient population.

1.6.2.1 5-FU AND CAPECITABINE

The anti-tumour effect of 5-FU mediated by TS inhibition as well as its effect on RNA is well established[204], but incorporation of 5-FU derivatives into DNA can also lead to cytotoxicity, but this may require an intact MMR system to detect the incorporated FdUTP[196].

Figure 1-7 5-FU metabolic pathway and mechanism of action



5′DFCR, 5′deoxy-5-fluorocytidine; 5′DFUR, 3′deoxy-5-fluorouridine; 5-FUR, 5-fluorouridine; CDD, cytosine deaminase; CES, carboxylesterase; DHP, dihydropyrimidinase; DPD, dihydropyrimidine dehydrogenase; FBAL, fluoro-b-alanine; FUH_2, dihydro-5-fluorouracil; MTHFR, methylenetetrahydrofolate reductase; OPRT, uridine monophosphate synthetase; RNR, ribonucleotide reductase; TK, thymidine kinase; TP, thymidine phosphorylase; UK, uridine-cytidine kinase 2; UP, uridine phosphorylase 1. Picture modified with permission from Walther *et al.*[1]

Variation in the enzymes mediating the incorporation into RNA and DNA, the conversion of the oral pro-drug capecitabine to 5-FU, or metabolism to inactive breakdown products can alter the intracellular 5-FU concentrations and cytotoxicity, leading to altered anti-tumour activity or systemic toxicity. For example, expression levels of TS are associated with drug efficacy: CRC with high levels of TS appear to have a poorer overall survival than tumours with low TS expression[205]. In many instances, it is

not clear what drives the changes in expression levels, although it is likely that underlying germline genetic variation or change is responsible in a proportion of the observed variation.

1.6.2.1.1 Thymidylate synthetase

Thymidylate synthetase (TS) is thought to be the dominant target for the active principle of 5-FU, fluorodeoxyuridine monophospate (5-FdUMP). A meta-analysis of TS expression suggested that higher expression of TS is associated with poorer overall survival (OS)[205]. Two main genetic determinants of TS expression have been described: a variable number tandem repeat (VNTR) polymorphism in the *TS* promoter-enhancer region (TSE) and a 6bp insertion/deletion polymorphism in the 3'-UTR of *TS*. The tandem repeat sequence is 28bp long, and usually present as either 2 or 3 repeats (TSER), although four and more repeats have been described[206]. An increase in the number of repeats leads to increased efficiency of mRNA translation and expression[207]. In addition, the G allele of a G/C SNP in the second repeat of the TSER3 correlates with even higher mRNA expression via conservation of an upstream stimulatory factor (USF) binding site in which the SNP lies[208]. The 6bp deletion polymorphism in the 3'-UTR region alters mRNA stability[209] and is associated with low mRNA expression[210]. The 'high expressing' variants have been associated with decreased survival in patients treated with 5-FU[206,211], and these three markers in combination may predict those with increased risk of recurrence in stage 2 and stage 3 colon cancer[212].

1.6.2.1.2 Dehydropyrimidine dehydrogenase

More than 80% of 5-FU is catabolised by dehydropyrimidine dehydrogenase (DPD), but levels of activity can vary widely between individuals: 3-5% of the population are partially DPD deficient, but complete deficiency is rare[213]. The bulk of DPD deficiency can be explained by more than 30 polymorphisms[214], which can lead to severe, sometimes life threatening toxicity after 5-FU treatment[215], although pre-treatment levels of DPD activity do not correlate well with toxicity[216]. It is impractical to screen all cancer patients for 30 polymorphisms *a priori*, and the lower toxicity associated with modern infusional or oral 5-FU based regimens make this less important. Despite extensive investigation, the pharmacogenetic basis of varied DPD activity remains to be fully elucidated, although advances in high throughput sequencing techniques may make the pre-treatment prediction of 5-FU toxicity achievable[217].

1.6.2.1.3 Methylenetetrahydrofolate reductase

Reduced methylenetetrahydrofolate reductase (MTHFR) activity creates variation in folate pools, indirectly increasing sensitivity to 5-FU. Two common polymorphisms in *MTHFR* lower enzyme activity[218]: the C677T polymorphism leads to a change of alanine to valine at position 222 and A1298C results in a glutamine to alanine change at position 429. Increased response to 5-FU treatment has been associated with the 677T allele[219,220] and to a lesser extent the 1298C allele. Recent studies suggest these polymorphisms affect capecitabine toxicity[221], as well as efficacy of 5-FU[222]. However, clinical data do not unequivocally support the influence of *MTHFR* genotype on 5-FU responsiveness, toxicity and patient outcome[223].

1.6.2.2 OXALIPLATIN

There is some evidence to suggest that genetic polymorphisms in detoxifying enzymes and DNA repair genes play an important role in treatment response to the DNA binding agent oxaliplatin. Decreased sensitivity to platinum agents has been attributed to diminished cellular drug accumulation; increased intracellular drug detoxification and enhanced DNA repair[224].

Glutathione-S-transferases (GSTs) are a class of phase II detoxification enzymes that target a wide variety of drugs for excretion by conjugation with glutathione. A number of isoenzymes and polymorphisms within these exist, with varying specificity, activity and tissue localisation[225]. The isoenzyme *GSTP1* is the primary enzyme for the detoxification of platinum derivatives. Two missense polymorphisms in *GSTP1*, I105V and A114V lead to decreased GSTP1 activity and predict neuropathy after oxaliplatin treatment[226]. There is no evidence for other isoenzymes or alleles (including null alleles) being associated with prognosis.

The primary anti-tumour mechanism of platinum derivatives is formation of DNA adducts, which interfere with DNA replication and require the activity of DNA repair enzymes to avoid cell death. Several polymorphisms in different DNA repair enzymes have been shown to correlate with function[227], however, association studies with outcome appear to be regimen and cancer-type specific. Out of six commonly studied functional polymorphisms in four repair genes (*ERCC1, ERCC2, XRCC1, XRCC3*) only *ERCC1* N118N and *ERCC2* K751Q were associated with overall survival in colorectal cancer[228], but not in all reports[229].

1.6.2.3 IRINOTECAN

The active metabolite of the topoisomerase I inhibitor irinotecan, SN-38, is conjugated and detoxified primarily by UDP-glucuronosyltransferase (*UGT1A1*). The number of TA repeats in the TATA element in *UGT1A1* correlates with reduced enzyme expression and

activity[230]. Individuals who are homozygous for the 7-repeat, also known as the *UGT1A1**28 allele, have decreased degradation and clearance of SN-38, commonly leading to dose-limiting neutropaenia[231]. Initially, several groups reported this association but findings were heterogeneous in significance and effect size. A recent meta-analysis established that the incidence of toxicity in *UGT1A1**28 patients was positively correlated with dose used[232]. A commercial genetic test was approved by the FDA in 2005 to make dosing recommendations based on *UGT1A1* genotypic results and avoid life-threatening neutropaenia, a first for any chemotherapy agent. The routine use of this test by oncologists has been limited in clinical practice because with the lower irinotecan doses used in combination regimens with 5-FU, haematological toxicity has decreased. The prognostic impact of *UGT1A1**28 has not been established[233].

At present, there are no pharmacogenetic markers which are useful in clinical practice, and those with a published positive association still require further validation.

1.6.3 Unbiased high-throughput screening

High-throughput arrays for evaluating mRNA expression levels or SNPs have opened new avenues of marker discovery by moving from hypothesis driven, targeted research to unbiased screening of the whole genetic spectrum. The former is aimed at quantitative markers, while DNA based techniques analyse discrete markers. The bulk of biomarker research has focussed on somatic changes, but germline variation equally can have an impact on prognosis and response to therapy, although to date only the few pharmacogenetic markers in the previous section have been described. For prognostic purposes, gene expression analysis is classically performed on tumour tissue, although comparative analyses with normal surrounding mucosa have been published, while genome-wide DNA screens have focussed on the germline, particularly in the search for disease susceptibility loci.

1.6.3.1 GENE EXPRESSION SIGNATURES

Gene expression analysis holds great promise for the understanding of functional differences between tumour and normal tissue, and with it the hope of developing meaningful signatures that stratify patients further beyond pathological staging. In CRC, much effort has been expended to repeat the perceived success in breast cancer, both by analysing tumour tissue[234-242] or the surrounding mucosa[243-246]. Similar analysis has also been applied to the study of miRNA arrays in tumour tissue[247-249].

The first published signature in CRC by Wang *et al.*, identified a 23-gene prognostic signature based on 31 relapses in 74 stage 2 patients[250]. This was subsequently subjected to validation both independently and by the same group. Based on 25 relapses in 50 stage 2 patients, the former found a positive predictive value of 67%, only marginally better

than the toss of a coin and therefore proposed a separate set of 30 genes[238], while the latter found that using only 7 genes of the original set performed better in their validation set[241], this time based on 123 patients and further validated in 104 patients. In common with other studies describing signatures based on differentially expressed genes (range 7 to 72), the overlap between the 30 and 7-gene signatures, even when in the same pathological stage, has been poor. One of the more convincing studies concerned miRNA expression, with 74 patients of all stages in the testing set, and 110 patients in the validation set, which suggested a single miRNA (miR21) as a prognostic marker. Again, this has not yet been validated further.

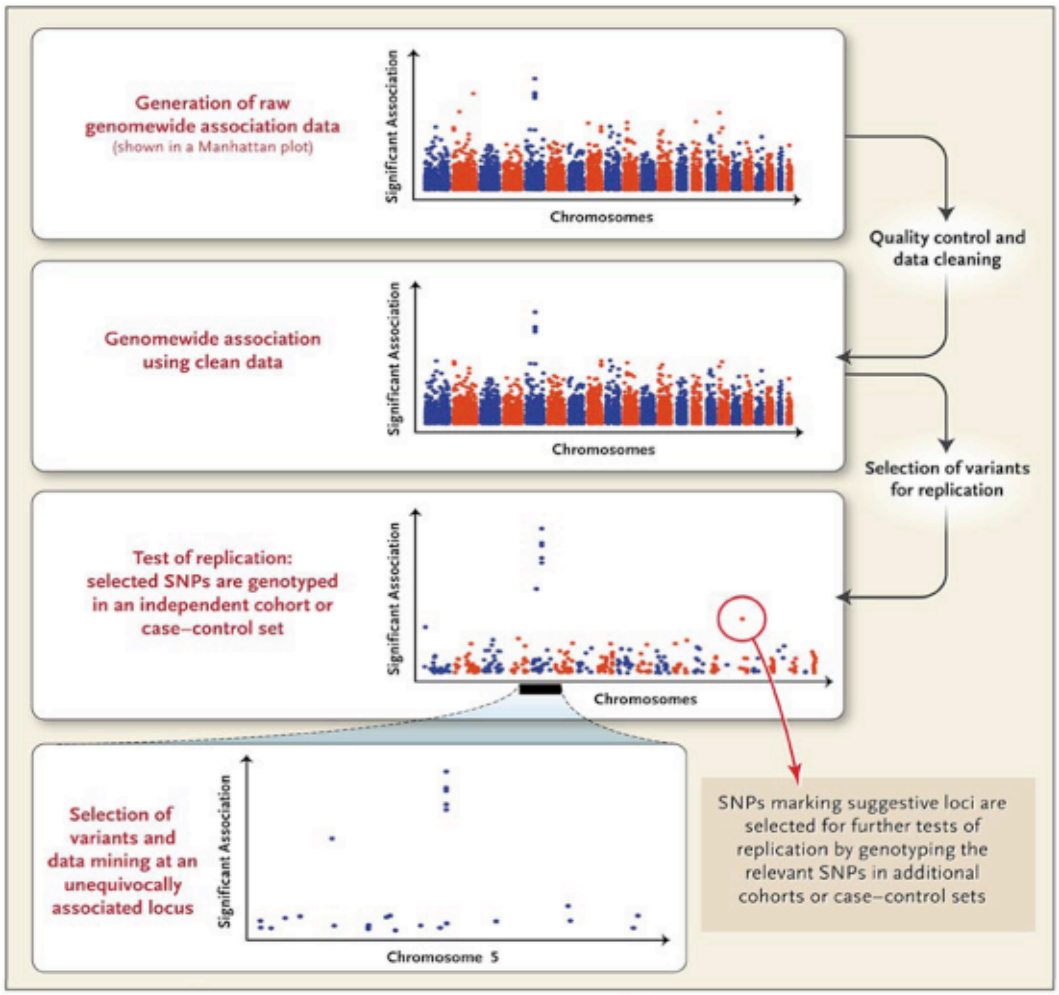
The reasons for this include that studies were underpowered, looking at sample sets as small as 20 patients while analysing the expression of tens of thousands of genes; validation was not always performed; and correction for other prognostic variables was omitted from the main analysis, leading to over-fitting of the prognostic model with poor reproducibility and a high false discovery rate[251]. This problem is not unique to CRC, nor has the quality of studies improved with increased familiarity with the technology[252-254] and even for the first signature, described in breast cancer, concerns have been raised over the inclusion of the training set in the initial validation set[255], and the insufficient patient numbers required to derive a robust signature. Many different signatures could be derived from the same data depending on the clustering parameters[256], and a sample set of at least 3000 patients may be necessary to identify a unique, robust signature[257]. Efforts are underway to generate a gene expression signature from fixed paraffin embedded tumour tissue using quantitative Real-Time PCR[258]. The initial signature was generated in a set of over 500 samples and will be validated in a large (n=2,000) cohort of colon cancer patients enrolled in the QUASAR trial[137].

Meta-analysis is commonly performed for other markers combining small datasets to boost numbers, but this is difficult and unreliable for expression signatures due to the differing methodologies applied in each study[259], leading to a high false positive and negative discovery rate[260]. Given the undoubted potential — and it is important to state that even a signature which does separate relapse from non-relapse only imperfectly may add to the clinical decision making process — there must nonetheless be a concerted effort to find sample sets of sufficient size to allow the definition of a robust, reproducible signature in CRC as "*small sample sizes might actually hinder the identification of truly important genes*"[261], and the emphasis on the detection of bias and chance and the validation of the findings remain as important as ever[262].

1.6.3.2 GENOME-WIDE ASSOCIATION STUDIES (GWAS)

Genome-wide association studies using high-density SNP arrays have been successful in the identification of susceptibility loci for colorectal[42,43], prostate[38] and breast cancer[39], non-malignant disease[40,41], and physiological variation[263,264]. They have relied on the premise that many diseases have a heritable component, and that traditional linkage analysis using a collection of affected families will only identify rare, high-penetrance susceptibility alleles. GWAS attempts to define common alleles that confer only a moderate risk for the disease in question (see section 1.3.2.2).

Figure 1-8 Basic structure of a GWAS



Initially all genotyped SNPs are analysed for an association with the complex trait in question. SNPs that have failed, are non-informative and those that appear to be unreliable are excluded before a selection is made to take a subset forward into verification in other cohorts. Further fine-mapping is (usually) required to determine the causative variant. Picture reproduced with permission from Hardy *et al.*[265] (modified).

It is likely that 'host' factors — the genetic make-up of normal tissue — can have an influence on outcome, either as prognostic or predictive marker. Genetic variation associated with genes involved in cell adhesion and motility may make early metastasis more likely, and cell cycle checkpoint associated variation may make the cell more tolerant of genomic abnormality and thus more resistant to chemotherapy.

Pharmacogenomic variation is likely to play a role both in host response (toxicity) and tumour response to chemotherapy. This is discussed above, but GWAS has the potential to uncover new variants affecting the metabolism and efficacy of anti-cancer agents.

Until all genetic variation has been described and can be tested for simultaneously, GWAS necessarily relies on the association of a tested marker with the true or 'causal' variant affecting outcome, although there is no inherent reason why the tested marker could not be the true variant. This means that all identified markers meeting a specified significance threshold need to be followed up irrespective of biological rationale, although if the marker identified is not in complete linkage disequilibrium with the true variant, replication would be more difficult. Once identified, however, routine clinical use would be easy as the necessary DNA could easily be extracted from a standard blood sample and genotyping is robust (Figure 1-8).

It is conceptually possible to apply the same techniques to tumour DNA, although prognostic markers may be due to mutations within the malignant tissue and not be captured by the SNPs analysed if these do not predispose to the mutation in the first place, nor would this capture epigenetic events which undoubtedly play a role in tumorigenesis and could also affect outcome. In malignant tissue, however, SNP typing platforms would give information about copy number changes (deletions, insertions, loss of homozygosity) that can affect gene expression (functional aneuploidy[266]).

To date, no GWAS looking at outcome, either in untreated or treated patients has been published for any cancer. The main obstacle is that for the analysis of several hundred thousand SNPs, in order to retain a degree of statistical power, sample sets need to be large[37]. While it is relatively easy to collect these sets for the analysis of risk, outcome, if nothing else, requires follow-up for at least three years to assess progression-free survival (DFS) with top quality data, in a clinical trial setting if at all possible.

1.7 The germline as the driving factor in CRC?

The encouraging improvement in patient outcome over the last 20 years has been followed by a plethora of markers of prognosis and response to anti-cancer therapy, the majority of which fail to demonstrate clinical utility. Many more DNA and RNA-based markers of prognosis than those described in section 1.6 have already been described, often only once and in small series. They are all based on somatic abnormalities detected in CRC cells, and fall into the category of hypothesis driven markers. Given that to date the only (predictive) marker to have been described this way with sufficient evidence to justify routine clinical assessment is *KRAS* mutational analysis for the selection of anti-EGFR strategies, new approaches to marker discovery are required. It is reasonable to assume that unbiased screening of either DNA or RNA based markers will significantly

advance our understanding of cancer biology, and allow for quasi individualised treatment to be given, exploiting inter-individual CRC differences to maximise treatment benefit.

The success of GWAS in defining novel CRC susceptibility loci shows that germline DNA variability confers CRC risk. Given that tumour cells arise from normal colonic mucosa, they start off with the same germline DNA, but acquire somatic mutations in a few key regulators and checkpoints. Anything that makes these mutations more likely to occur will increase CRC risk, and at the same time, any variation at the cellular level hardwired into the blueprint of the cell, namely DNA, will be present in tumour as well as normal cells.

At present, up to 35% of CRC is thought to have a familial component[16], but this may only be a lower limit estimate[267], with many more CRC attributable to genetic causes through gene-environment interaction, genetic variation modifying the impact of the environmental factor in the individual. Even the well-described hyper-methylation of the *MLH1* promoter in sporadic MSI+ve CRC[106] could be determined by variation in genes governing levels of DNA methylation.

By analogous logic, outcome in CRC as a complex trait may also be influenced by germline genetic variation in a large proportion of cases, predisposing to somatic changes associated with prognosis, as well as altering the interaction between tumour and its environment, and the overall host response to tumour and treatment. It is difficult to predict the functional consequences of any novel marker that may be identified by such an approach - hypothesis driven research has not yielded any validated candidates - but conceptually they should fall into one of three categories: markers that alter the likelihood of early and frequent metastasis, thus determining whether surgery could be curative; markers that make cancer cells more tolerant of its sub-optimal environment (hypoxia, damage induced by chemotherapy, host response to the tumour) thus conferring a growth advantage; and markers that allow the cancerous cell to develop resistance to any chemotherapy given. The last category could include changes that alter the host response to treatment, meaning that it may not be the malignant cells that are more tolerant to chemotherapeutic agents but that, for example, the more efficient degradation of these agents leads to the delivery of a sub-lethal dose to the cancer.

In the absence of validated prognostic factors outside pathological staging, however, the genetic factors governing the 'prognostic process' are unknown, and an unbiased, high-throughput approach is most likely to provide insights into the effect of the germline on CRC prognosis. This thesis describes a search for novel markers of prognosis and risk using GWAS, and the attempted validation of several hypothesis driven markers.

Chapter 2

Methods

2.1 Samples used

The genotype data used in this thesis was derived from 13 cohorts, this was generated by collaborators for multiple loci for the following cohorts: 1958 Birth Cohort, PETACC, London Phase 2, Scotland Phase 1, and Scotland Phase 2. rs6983267 was typed by collaborators for these cohorts: Quasar 1, Australia, and Epicolon. For Corgi 2 (London Phase 1), Denmark, and Finland, some genotypes were generated by collaborators, others by Axel Walther. (Almost) all genotypes for VICTOR and all for Quasar 2 were generated by Axel Walther (Table 2-1).

All cases had pathologically proven adenocarcinoma of the colon or rectum. All participants gave informed consent for the collection of blood and tumour samples, and clinical and pathological information. All studies had local ethical review board approval in accordance with the Declaration of Helsinki. DNA was extracted from samples using conventional methodologies and quantified using PicoGreen (Invitrogen, Paisley, UK) or Nanodrop 1000 (Thermo Scientific, Waltham, MA).

2.1.1 VICTOR

Germline DNA was available for 947 patients enrolled in the VICTOR trial (612 males, 335 females). All patients had stage 2 (480 patients) or 3 (467 patients) CRC treated with surgery, and received adjuvant 5-FU based chemo- and/or (neo-)adjuvant radiotherapy as appropriate for stage and site of disease. This was at the discretion of the treating physician. Patients were subsequently randomised to receive rofecoxib or placebo for 2 or 5 years. The trial was stopped because of concerns over the cardiac side-effects of rofecoxib. Because of early closure, only 30 patients received trial medication for more than 24 months (median 8 months, range 0 - 28 months).

Table 2-1 Responsibility for genotyping in the cohorts utilised

Cohort	Genotypes	Performed by
1958 Birth Cohort	all	Collaborators
Australia	rs6983267	Collaborators
Corgi 2	all loci in Chapter 7	Axel Walther
	all others	Collaborators
London Phase 2	all	Collaborators
Denmark	all	Axel Walther
Epicolon	rs6983267	Collaborators
Finland	rs6687758	Axel Walther
	rs6983267	Collaborators
PETACC 3	all	Collaborators
Quasar 1	rs6983267	Collaborators
Quasar 2	all	Axel Walther
Scotland Phase 1	all	Collaborators
Scotland Phase 2	all	Collaborators
VICTOR	all except	Axel Walther
	rs6983267 in FFPE samples	Collaborators
	rs10411210	Collaborators
	rs12953717	Collaborators

Collaborators ultimately provided all genotypes, either directly, or by providing DNA for typing in the laboratory of Prof Ian Tomlinson. I am grateful to Dr Enric Domingo for helping with the genotyping of a subset of the VICTOR cohort.

Germline DNA was extracted using from blood using standard methods by the Department of Pharmacology at Oxford University. The data cut-off for the initial analysis presented in chapter 3 was 17 November 2007, and the latest available follow-up was 29 April 2009. Genotyping was performed on HumanHap300 and HumanHap370 arrays, and those who had relapsed were typed first to enrich for cases should the funding for the arrays become an issue at a later stage, therefore there was an excess of stage 2 patients among those typed on the Hap370 arrays (p=2.12e-05). 526 patients were typed using Hap300 arrays , the remainder on Hap370 arrays. Individual SNPs were typed using KASPar allele specific PCR.

184 additional patients were typed for rs6983267 from FFPE tissue, of these, 10 patients were censored at date of death from other causes, and five were excluded as follow-up data were missing.

2.1.2 PETACC

Germline DNA from FFPE tissue of patients enrolled in the PETACC 3 trial was available at the University Hospital Gasthuisberg, Leuven, Belgium. Patients had stage 2 (375 patients) or 3 (841 patients) colon cancer, and were randomised to receive either 5-FU alone or in combination with irinotecan following curative resection. Individual SNPs were typed using a fluorogenic 5' nuclease assay (TaqMan, Applied Biosystems, CA),

multiple SNPs were typed using MassARRAY® iPLEX Gold single-base extension technology (Sequenom, CA). 4-year DFS data was mature in March 2008.

2.1.3 GWAS for CRC risk loci

The London phase 1 GWAS was conducted using the Illumina HumanHap550 BeadChips, and the Scottish phase 1 GWAS was conducted using various Illumina SNPs arrays. In London and Edinburgh phase 2, genotyping was conducted using Illumina Infinium custom arrays. The 1958 Birth Cohort was typed on HumanHap550 BeadChips. All genotyping was performed in accordance with Illumina's protocols. No survival data was available except for Scotland Phase 1.

2.1.3.1 LONDON PHASE 1 (COLORECTAL TUMOUR GENE IDENTIFICATION 2, CORGI 2)

Phase 1 comprised 940 cases with colorectal neoplasia (443 males, 497 females) ascertained through the Corgi consortium. All had at least one first-degree relative affected by CRC and one or more of the following phenotypes: CRC at age 75 or less; any colorectal adenoma at age 45 or less; three or more colorectal adenomas at age 75 or less; or a large (>1 cm diameter) or aggressive (villous and/or severely dysplastic) adenoma at age 75 or less. Controls (n=965; 439 males, 526 females) were spouses or partners unaffected by cancer and without a personal family history (to second-degree relative level) of colorectal neoplasia. All cases and controls were of European ancestry and from the UK[46].

2.1.3.2 LONDON PHASE 2 (PHASE 2)

Phase 2 consisted of 2,873 CRC cases (1,199 males, 1,674 females, mean age at diagnosis 59.3 years; SD ± 8.7) ascertained through two ongoing initiatives at the Institute of Cancer Research/Royal Marsden Hospital NHS Trust (from 1999 onwards: the NSCCG27 and the Royal Marsden Hospital NHS Trust/Institute of Cancer Research Family History and DNA Registry. A total of 2,871 healthy individuals were recruited as part of ongoing National Cancer Research Network genetic epidemiological studies, NSCCG (n=1,235), the Genetic Lung Cancer Predisposition Study (1999–2004; n=917) and the Royal Marsden Hospital NHS Trust/Institute of Cancer Research Family History and DNA Registry (1999–2004; n=719). These controls (1,164 males, 1,707 females; mean age 59.8 years; SD ± 10.8) were the spouses or unrelated friends of individuals with malignancies. None had a personal history of malignancy at time of ascertainment. All cases and controls were British of recent European ancestry, and there were no obvious differences in the demography of cases and controls in terms of place of residence within the UK[46].

2.1.3.3 SCOTLAND PHASE 1

Phase 1 included 1,012 CRC cases (518 males, 494 females; mean age at diagnosis 49.6 years; SD ± 6.1) and 1,012 age- and gender-matched cancer-free population controls (518 males, 494 females; mean age 51.0 years; SD ± 5.9). Cases were enriched for genetic aetiology by early age at onset (age less than or equal to 55 years). Known dominant polyposis syndromes, hereditary non-polyposis colorectal carcinoma/Lynch syndrome or bi-allelic *MYH* mutation carriers were excluded. Control subjects were population controls, matched by age (± 5 years), gender and area of residence within Scotland[46].

2.1.3.4 SCOTLAND PHASE 2

Phase 2 comprised 2,057 CRC cases (1,249 males, 808 females; mean age at diagnosis 65.8 years; SD ± 8.4) and 2,111 population controls (1,257 males, 854 females; mean age 67.9 years; SD ± 9.0) ascertained in Scotland. Cases were taken from an independent, prospective, incident CRC case series and aged <80 years at diagnosis. Control subjects were population controls matched by age (± 5 years), gender and area of residence within Scotland[46].

2.1.3.5 WELLCOME TRUST CASE CONTROL CONSORTIUM 1958 BIRTH COHORT (58C)

This cohort was from the collection of DNA from all births in England, Wales and Scotland during a single week in 1958, consisting of 1,438 population controls from the Wellcome Trust Case Control Consortium 1958 birth cohort (58C, also known as the National Child Development Study) [46,268].

2.1.4 Finland

Germline DNA from blood was available for 1055 patients with CRC, and 864 controls (randomly selected anonymous Finnish blood donors), all ascertained in south-eastern Finland, the Finnish Colorectal Cancer Predisposition Study (FCCPS). DNA was extracted using standard methods, and whole-genome amplified by KBioscience, Hoddesdon, UK. 5-year OS was mature in August 2003. Genotyping was performed by allele specific PCR.

2.1.5 Denmark

Germline DNA was available for 599 patients with CRC at the Department of Molecular Medicine, Aarhus University Hospital, collected between February 1999 and November 2007. Informed consent was obtained from all patients according to local ethical regulations and the study was approved by the local ethical committee according to the Helsinki Declaration. All patients were of white Caucasian background. Genotyping was performed by allele specific PCR.

2.1.6 Quasar 1

Limited germline DNA was available for 492 patients enrolled in the Quasar 1 trial of adjuvant 5-FU chemotherapy versus observation for stage 2 colon cancer[137]. DNA was extracted from paraffin embedded tissue. Median follow-up was 5.4 years and analysis was truncated at 8 years, beyond which there were no events. Genotyping was performed by allele specific PCR.

Table 2-2 Summary of cohorts used

				Stage			
Cohort	Number	CRC	CRC survival	1	2	3	4
58BC	1436	Controls	n/a	-	-	-	-
Australia	477	Cases	2-year DFS	43	73	243	118
Corgi 2	1905	Both	none				
London Phase 2	5719	Both	none				
Danish	599	Cases	2-year OS	81	235	186	81
Epicolon	539	Cases	4-year DFS	68	230	146	95
Finland	1,045	Both	5-year OS	200	390	272	126
PETACC	1,216	Cases	4-year DFS	-	375	841	-
Quasar 1	492	Cases	5-year DFS	-	492	-	-
Quasar 2	494	Cases	1-year DFS	-	163	331	-
Scotland Phase 1	1982	Both					
Scotland Phase 2	4116	Both	n/a	-	-	-	-
VICTOR	1,136	Cases	3-year DFS	-	575	561	-

2.1.7 Quasar 2

Germline DNA was available for 539 patients enrolled in the Quasar 2 trial of capecitabine with or without bevacizumab for stage 2 (163 patients) and 3 (331 patients) colon cancer. Toxicity data was available for each patient for the 10 categories specified in the trial protocol. The latest available follow-up was 29 January 2009. Genotyping was performed on HumanHap370 arrays.

2.1.8 Australia

Tumour DNA from FFPE was available for SNP rs6983267 analysis from 477 patients with colorectal cancer (367 colon, 110 rectum) of all stages treated at the Royal Melbourne Hospital, Western Hospital between July 1990 and July 2006. All patients were selected from a prospectively collected comprehensive colorectal cancer database, Biogrid Australia[269]. 20 patients were missing follow-up information and were excluded. Written informed consent was obtained from all study subjects. A minimum of two-year follow-up was available for all cases. Genotyping was performed by allele specific PCR.

2.1.9 Epicolon

Germline DNA for 515 CRC cases (305 males, 210 females; mean age at diagnosis 70.6 years; SD ± 11.3) and 515 controls (290 males, 225 females; mean age 69.8 years; SD ± 11.7) was ascertained through the EPICOLON initiative, a prospective, multicenter, nationwide study aimed at compiling prominent epidemiological and clinical data with respect to hereditary non-polyposis colorectal cancer and other familial colorectal cancer forms in Spain. This cohort consists of an incident series collected in Barcelona[42]. Genotyping was performed by allele specific PCR.

2.1.10 VQ58

The VICTOR, Quasar 2 and 1958 Birth cohorts were combined as a single group for the analysis of the impact of rs6687758 on CRC risk in Chapter 7. It does not contain any patients not already mentioned above, and has been used in the discovery of other CRC risk loci[42].

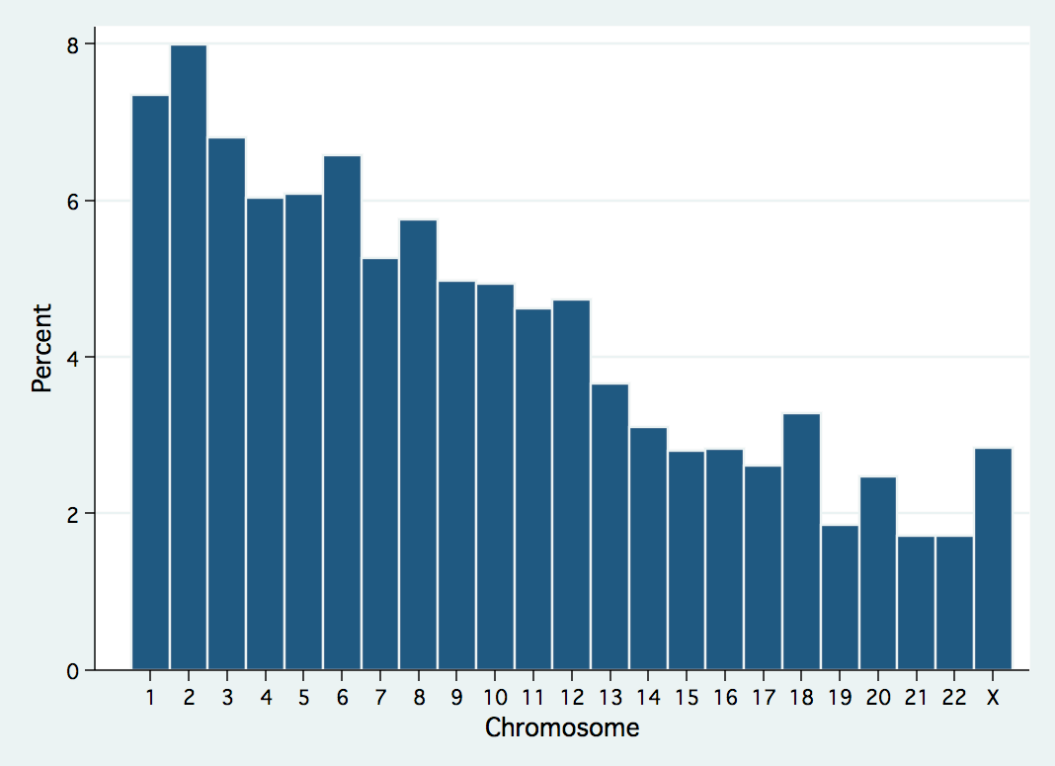
2.2 Experimental methods

2.2.1 High-throughput genotyping: Illumina Infinium arrays

High-throughput genotyping was performed on Illumina HumanHap300 or HumanHap370 (Hap300 and Hap370 for short) arrays (BeadChips), containing 318237 tagSNPs (Hap300) or 318237 tagSNPs plus 52167 markers in copy-number variation (CNV) regions on the Hap370 arrays (370404 markers).

All arrays were 'Duo', meaning that two samples could be processed on each array. Figure 2-1 gives the distribution of SNPs on the Hap300 arrays by chromosome. The decreasing number of SNPs with increasing chromosome number reflects that higher numbered chromosomes are smaller. The median density of SNPs per chromosome is 108 SNPs per Mb, ranging from 89 SNP per Mb (chromosome 15) to 137 SNPs per Mb (chromosome18). Both types are based on the Infinium II chemistry, proprietary to Illumina (San Diego, CA). Several multimer sequences are attached to beads, each consisting of 50 bases for target sequence recognition, and 22 bases for determining positional information (the IllumiCode). Each bead carries tags for only one SNP, and each bead type is replicated approximately 15-17 times per Duo array. The beads are randomly distributed on the array, and the positional tag is used to determine the position of each bead, and with it the associated SNP, prior to use with sample DNA. This positional information is required when scanning the arrays, as otherwise there will be no correlation between dye intensity and SNP analysed.

Figure 2-1 Distribution of Hap300 SNPs by chromosome



Percentage of total Hap300 content per chromosome. The decreasing percentage of content going right reflects the smaller size of these chromosomes. The X-chromosome is included in this figure for illustration purposes only, these SNPs were not analysed in this thesis.

All alleles are coded A and B according to the Illumina convention (Table 2-3): the A allele of any SNP is always the adenosine (A) or thymine (T), the B allele always the cytosine (C) or guanine (G) . All SNPs on the array are either A or T substituted by either C or G, there are no A/T or C/G SNPs present (see also section 2.3.1).

The genotyping is based on a single base extension assay, with one dye for the A allele, and another for the B allele. After annealing of the sample DNA to the template sequence tag on the bead, the single base extension step adds one hapten labelled base to the distal, 3' end of the bead sequence, complementary to the sample. This hapten signal recognised by a two-colour antibody-based staining and signal amplification step (red dinitrophenol for A, green biotin for B)[270].

Table 2-3 Illumina allele naming convention

Actual Base	Illumina name
Adenosine (A) or Thymine (T)	A
Cytosine (C) or Guanine (G)	B

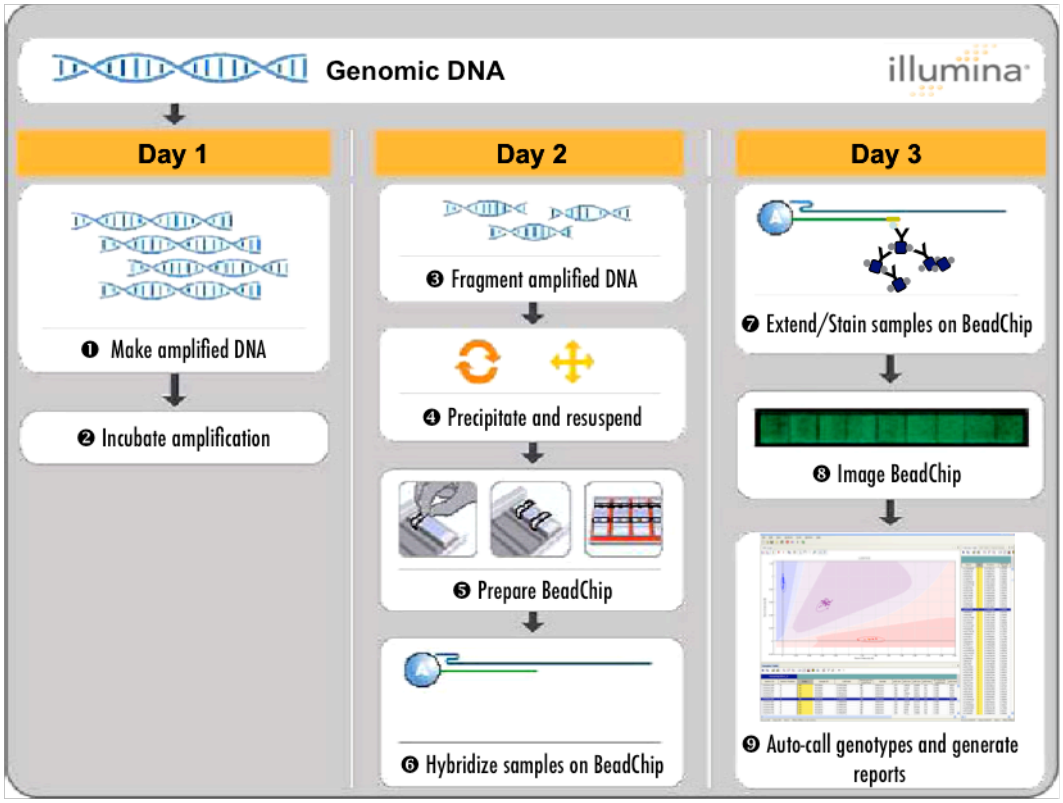
The Illumina naming convention allocates the same name to a particular allele of a SNP, irrespective of whether the SNP is typed on the +strand or -strand, as A on the +strand would be paired with T on the -strand. Likewise, C would be paired with G, and because no A/T or C/G SNPs are on the array, this convention allows all SNPs to be unambiguously represented by A and B.

2.2.1.1 METHOD

All DNA was and freeze-dried and resuspended in 15μL dH_2O at 50ng/μL (total 750ng). The genotyping was performed according to Illumina's protocols. The workflow comprises of neutralisation and denaturing for whole-genome amplification of the sample DNA. The amplified DNA is fragmented using 'end-point fragmentation' to avoid over-fragmentation. The fragmented DNA is collected by precipitation, resuspended and applied to the Illumina BeadChips followed by overnight incubation.

The next day, excess or non-specifically hybridised DNA is washed off, followed by single base extension with incorporation of detectable labels that are stained with antibody. Finally, the arrays are coated with a proprietary polymer and scanned on the Illumina BeadStation (Figure 2-2).

Figure 2-2 Illumina workflow for BeadChip analysis



Per sample workflow involves three day protocol with overnight amplification of the DNA, followed by preparation of the DNA for incubation on the array, overnight from day 2. On the third day, the single-base extension step is followed by staining and scanning of the finished array. In manual mode, batches of 16 samples are the most efficient way of processing. The single base-extension step occurs at the distal end of the multimer attached to the bead, after which the sample DNA is stripped off (figure courtesy of Illumina, Inc, CA, and modified with permission).

2.2.1.2 INTERNAL QUALITY CONTROLS

In addition to the beads on the arrays used for genotyping purposes, there are several Beadtypes for quality control purposes. These are sample independent controls to assess the performance of specific steps in the protocol (staining, single-base extension, target

removal, hybridisation), and sample dependent controls to assess the performance of the protocol across samples (stringency of hybridisation, non-specific binding, non-polymorphic controls). The controls assess one or the other of the two colour channels (red and green), or both. For each control, there are a number of Beadtypes with 15-17 replicates each. The internal controls allow exploration of reasons for failure or underperformance of samples and arrays[271].

2.2.1.2.1 Staining controls

Twenty Beadtypes for each colour are labelled only with varying levels of the hapten of that colour, testing the efficiency of the signal amplification step. It s independent of sample and hybridisation. This tests both the red and green channel.

2.2.1.2.2 Extension controls

Ten Beadtypes for each of the four bases. Each bead has hairpin multimers attached, which allow single-base extension without sample at the end of the hairpin. These beads should stain in the normal fashion, and test both colours (A and T for red, C and G for green).

2.2.1.2.3 Target removal controls

Ten Beadtypes with multimers that do not allow extension at the distal, 3' end. Target removal controls, short sequences of DNA complimentary to the target removal control bead sequences are added after the overnight hybridisation of sample, and after annealing, extend at the 3' end of the target removal control DNA. After the stripping of sample DNA, the target removal DNA sequences should also be removed, and no colour signal detectable from the target removal control beads. All beads have a green extension step.

2.2.1.2.4 Hybridisation controls

Hybridisation is tested at three concentration levels of synthetic template (0.2pM, 1pM, 5pM), with ten Beadtypes for each level. Synthetic template with perfect complimentarity is added after the overnight hybridisation of sample, and subjected to single-base extension. Colour intensity at each beads should be correlated to the concentration of the synthetic template. All beads have a green extension step.

2.2.1.2.5 Stringency controls

This is a sample dependent assessment of the stringency of hybridisation of non-polymorphic template DNA. The 20 Beadtypes have varying levels of mismatch in the multimer sequence, from perfect match to 12bp mismatch. With increasing mismatch, the staining intensity will reduce. All beads have a red extension step.

2.2.1.2.6 *Non-specific binding controls*

Forty Beadtypes with multimers complimentary to bacterial sequences, with a distal, 3' end available for single-base extension if sample DNA has annealed. Human DNA should not hybridise to these sequences under normal stringency conditions (as used in the Illumina process). This is monitored for both red and green channels.

2.2.1.2.7 *Non-polymorphic controls*

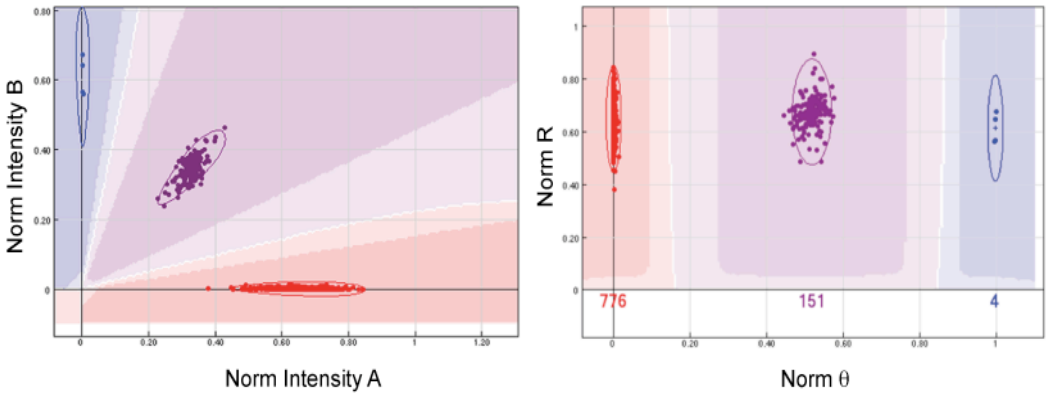
This tests the overall performance of the process, by using non-polymorphic template DNA. 10 Beadtypes are included for each base, and the colour intensity at these loci permits comparison of assay performance across samples.

2.2.1.3 GENOTYPING QUALITY CONTROL ALGORITHMS

2.2.1.3.1 *GenTrain score*

The GenTrain score is a SNP specific, statistical score, defining how well an individual SNP has assayed. It takes into account the shapes of the clusters and their relative distance to each other. The GenTrain scores can range from 0 to 1, with 0 representing poorly performing/failed genotyping, and scores close to 1 representing the best performing genotyping at that SNP. It is not a score routinely used by the Illumina BeadStudio software to assess the robustness of genotypes, although GenTrain score <0.3 is likely to reflect a poorly performing SNP.

Figure 2-3 Sample genotype plot generated in BeadStudio (rs1955049)



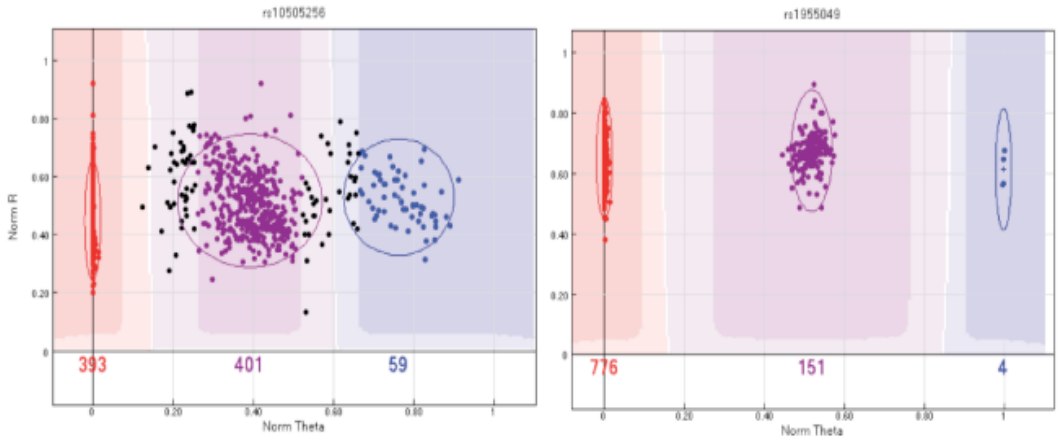
Plot of the normalised intensities of red and green channels for rs1955049, the SNP with the highest GenTrain score (left). The genotypes clearly fall into three clusters. The data points are converted into a plot of angle from the origin θ versus the log normalised intensity ($\theta=(2/\pi)$ x arctan(B/A) where B is the green channel (y-axis) and A the red channel (x-axis). This new plot (right) forms the basis of the genotype calling algorithm.

2.2.1.3.2 *GenCall score*

The GenCall score is a sample and SNP specific quality score that is based on cluster position and definition in the genotype cluster plot (Figure 2-3), and is a measure of the distance of the individual genotype position from the centre of the cluster in that plot. Its value is influenced by the signal intensity, dispersion of genotypes, overlap of clusters, and

angle θ of the cluster. GenCall scores are calculated for each genotype, and give an estimate of how reliable that genotype is (for a specific sample and SNP). It ranges from 0 to 1, with a lower score indicating increasing distance from the centre of a cluster, implicating worse clustering. The sensitive range of the GenCall score is in the region of 0.2-0.7 depending on the Illumina platform used[272]. Below that range, the SNP generally has failed, above that, and the SNP is generally robust. For Infinium arrays (all the arrays used in this thesis), the recommend GenCall cut-off is 0.15 due to the very tight clustering seen with this assay[273], i.e. a SNP may be legitimately associated with a cluster, but because all other samples have clustered so tightly, its GenCall score will be relatively lower.

Figure 2-4 Distinction between a successful and a failed SNP



The cluster separation between AB and BB is poor, and samples lying between the two clusters cannot be reliably allocated to either one or the other, and the SNP was excluded (left). A well performing SNP with good cluster separation, there is no ambiguity about the genotype calls (right).

2.2.1.3.3 Call frequency

Call frequency is a per SNP score giving the percentage of samples called for that SNP. It is the number of samples called divided by the total number of samples. It assesses how well a SNP has performed across samples, and is driven by GenTrain and GenCall score.

2.2.1.3.4 Call rate

The call rate is a per sample measure of the number of SNPs called. It is the number of SNPs successfully called divided by the total number of SNPs. It reflects the quality of the DNA, as well as the fidelity of the process. It is ultimately also driven by the quality and GenCall scores as these determine the number of SNPs called.

2.2.1.4 GENOTYPE CALLING

The Illumina genotype calling and visualisation software BeadStudio determines clusters of genotypes by the colour at each bead locus: red for A homozygote, green for B homozygote, and yellow for heterozygote loci. The colour intensity at each bead is

measured and converted into an angle θ versus log-normalised intensity. The signal at each SNP locus should cluster into positions corresponding to the three genotypes (Figure 2-3).

The first step is self-normalising of the array, taking into account normal variation in signal intensity, variation in background intensity between the two channels, and the possibility of 'cross-talk' between the channels. The normalising process aids the formation of clusters, which are recognised by Illumina proprietary algorithms. This generates the GenTrain score for each SNP.

Using the intensity data from each array, based on the GenTrain score, individual samples are assigned to a cluster. During this process the GenCall score is generated. As the algorithm requires a minimum of 100 samples to function properly, all samples were analysed jointly at the end. Illumina also advises manual editing of SNPs with poorly performing automated clustering[273], this was performed on SNPs with low call frequency (Figure 2-5 and section 2.2.1.5).

Any SNP that had a GenCall score of 0 allocated by the calling software, or that on visual inspection of the cluster, had a poor cluster separation was deemed to have failed. These SNPs were not considered for the verification phase.

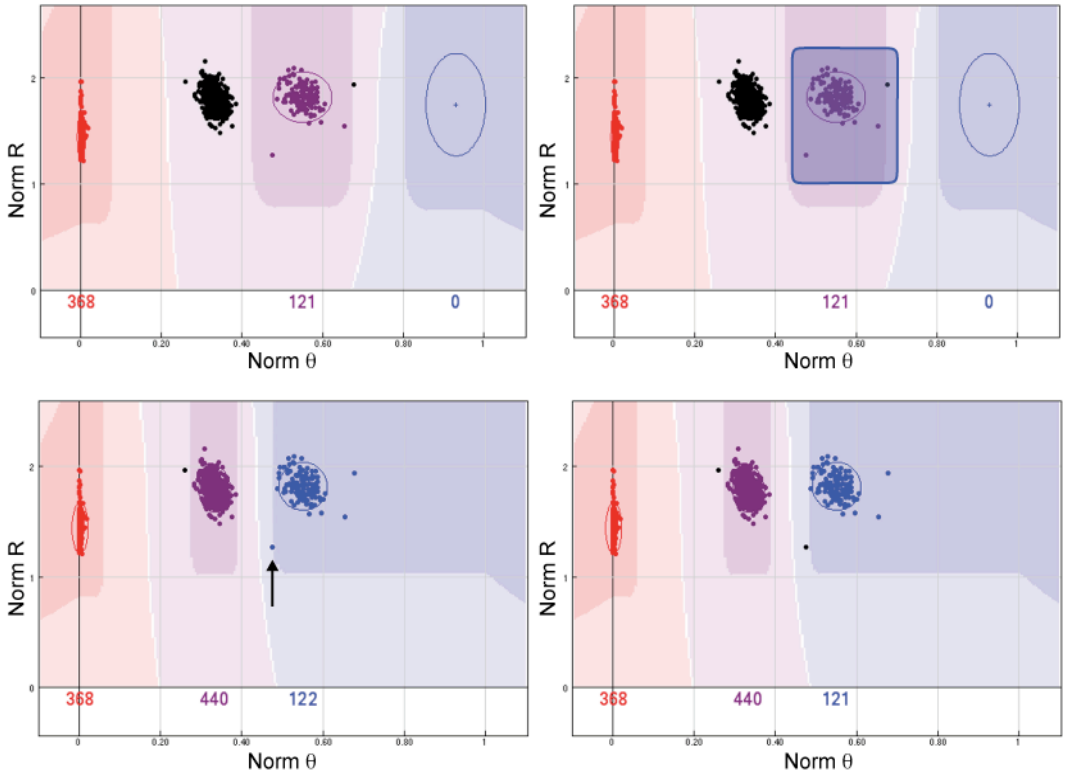
2.2.1.5 MANUAL RECALLING OF GENOTYPES

Poorly performing SNPs were inspected visually and, within the constraints of the calling software BeadStudio, reassigned genotypes based on the cluster distribution. The first step was to run the calling algorithm and then sort SNPs based on their call rates. The cluster plot for each SNP was then inspected starting with the least well performing SNPs, and if the automatic clustering was unsatisfactory, manual reclustering was attempted unless the SNP was poorly clustered (see Figure 2-4 left), when it was excluded.

Manual clustering within the BeadStudio software can be performed in two separate ways: firstly, selecting a cluster and forcing the algorithm to recognise this as a particular genotype, and secondly, by moving the cluster boundaries to include or exclude samples that are deemed to be included or excluded, respectively, in the cluster based on automated calling. In both instances, the genotype allocation is constrained by the algorithm so that it is not possible to allocate a sample any genotype but rather the plot position of the sample in relation to the other samples analysed is still taken into consideration. In Figure 2-5, both techniques are demonstrated: the rightmost cluster is not recognised as the BB cluster as it lies where normally an AB cluster would be expected (top left). By forcing the rightmost cluster to be recognised as BB, the other two clusters are unambiguously allocated the appropriate genotypes, but this now includes an

ambiguous sample, and moving the BB cluster area (blue ellipse) slightly to the right, will exclude this sample.

Figure 2-5 Impact of manual recalling



This SNP (rs10904494) clearly forms three clusters, but the calling algorithm determines that it does not conform to the expected locations of the clusters (the BB cluster should be centred around Norm θ=1) and does not call the AB cluster correctly (top left). By selecting the samples of the rightmost cluster, and forcing the calling software to use these as the BB cluster (top right, shaded), all three clusters are assigned appropriate genotypes at this locus (bottom left), even if the generated GenTrain score remains low (0.367). The sample indicated by the arrow does not clearly belong to either AB or BB cluster, and by shifting the cluster position, this can be excluded from the BB cluster (bottom right).

2.2.2 Single-marker genotyping

This section describes the standard conditions for each technique. All variations and primer sequences are given in the relevant appendices. All reagents are CRUK in-house versions, where this differs, it is indicated in the text. The manufacturer is given for all proprietary reagents. Primers were typically designed using Primer3 (Whitehead Institute of Biomedical Research) and obtained from SigmaAldrich (Gillingham, UK).

2.2.2.1 STANDARD PCR FOR GENESCAN AND SEQUENCING

The PCR for GeneScan and sequencing reactions is made up to 25µl total volume, and typically contains 30-50ng of template DNA per reaction. The Mg^{2+} concentration was 1.5mM, or adjusted during optimisation for each primer pair (usually 2.5mM), DMSO replaced Q-solution (Quiagen, Crawley, UK) for the amplification of GC rich regions (Table 2-4).

Table 2-4 Reagents for standard PCR (all volumes in μL)

			GC rich regions	
Reagent	1.5mM Mg^{2+}	2.5mM Mg^{2+}	1.5mM Mg^{2+}	2.5mM Mg^{2+}
DNA	2.0	2.0	2.0	2.0
Taq Polymerase	0.25	0.25	0.25	0.25
Forward primer [20mM]	0.25	0.25	0.25	0.25
Reverse primer [20mM]	0.25	0.25	0.25	0.25
MgCl [25mM]	1.5	2.5	1.5	2.5
Buffer	2.5	2.5	2.5	2.5
Nucleotide	2.0	2.0	2.0	2.0
Q solution	5.0	5.0	-	-
DMSO	-	-	2.5	2.5
dH_2O	11.25	10.25	12.75	11.75

Reagents were mixed for the total number of samples to be typed and plated onto 96-well plates (ABgene, Epsom, UK), followed by the addition of DNA to each well. Plates were sealed with Thermowell sealers (Corning) and immediately run on a Tetrad PCR machine (MJ Research, Waltham, MA). The cycling conditions are given in Table 2-5 with a standard annealing temperature of 55°C. All primers underwent optimisation for magnesium content, the use of DMSO and annealing temperature prior to experimental use.

PCR products were visualised by agarose gel electrophoresis and ethidium bromide staining (Section 2.2.2.1.1).

Table 2-5 Conditions for standard PCR

Temperature	Time	
95°C	4 minutes	Initial denaturing
94°C	1 minute	
55°C	1 minute	For 35 cycles
72°C	1 minute	
72°C	10 minutes	Final extension
15°C	Until plates removed from machine	

2.2.2.1.1 Agarose gel electrophoresis

Depending on the size of the PCR product, 1, 2 or 3% agarose gels were used to visualise PCR product. The required amount of agarose (Invitrogen) was dissolved in 1X TBE buffer (Cancer Research UK) and ethidium bromide was added to a final concentration of 0.25μg/ml. The gels were set, samples were loaded and the gel was run in an electrophoresis chamber typically at 120V for 20 minutes. The DNA was then visualised with a White/2UV Transilluminator (UV Products, Cambridge, UK).

2.2.2.1.2 GeneScan analysis

Following primary PCR, products were visualised by agarose gel electrophoresis and ethidium bromide staining (Section 2.2.2.1.1). The PCR product was diluted with dH2O according to the intensity of the band run on an agarose gel (between 1:1 and 1:150) and 3μL volume of final dilution plated onto barcoded optical plates (ABgene, Epsom, UK). The plates were analysed using an AB3100 genetic Analyzer (Applied Biosystems, Woolston, UK) with 7-10ms injection time.

2.2.2.1.3 Sequencing

Following primary PCR, 5μL of PCR product were incubated with 2μL ExoSAP-IT (GE Healthcare, Amersham, UK) to remove any excess reagents and incubated for 30 minutes (Table 2-6).

Table 2-6 ExoSAP reaction conditions

Temperature	Time
37°C	15 minutes
80°C	15 minutes
15°C	Until plates removed from machine

Product from this reaction was diluted according to the intensity of the band on an agarose gel (Section 2.2.2.1.1) from the primary PCR, and amplified in a 2nd PCR (Table 2-7), yielding fragments of varying length with a base specific dye in the last base of the fragment.

Table 2-7 Reagents for sequencing PCR

Reagent	Amount [μL]
PCR product	4.0
Primer [20mM]	1.0
BDT	8.0
dH_2O	7.0

BDT - Big Dye Terminator (Applied Biosystems, Woolston, UK)

Table 2-8 Conditions for sequencing PCR

Temperature	Time	
95	5	
96	.5	
50	.75	35
60	4	
4°C	Until plates removed from machine	

Following the 2nd PCR step, the total reaction volume was placed in a DyeEx spin column (or plate if 96 samples were processed simultaneously, Quiagen, Crawley, UK) to remove unincorporated dye terminators, and spun at 2400rpm for 3 minutes. The eluant was

placed into bar-coded optical plates (ABgene, Epsom, UK). The plates were analysed using an AB3730 DNA Analyzer (Applied Bioscience, Woolston, UK). The output was read visually in each case.

2.2.2.2 KASPAR

KASPar is an allele specific PCR based on allele specific dye incorporation. The dye intensity is measured and converted into a genotype call. The biochemistry is proprietary to KBioscience (Hoddesdon, UK). All forward primers are labelled and chosen using KBioscience PrimerPicker software.

Table 2-9 Reagents for KASPar PCR

Reagent	Amount [μL]
DNA	1.0
Common primer [100mM]	0.025
Forward primer [100mM]	0.010
Reverse primer [100mM]	0.010
Taq Polymerase	0.039
MgCl [50mM]	0.096
4x Reaction Buffer	3.0
dH_2O	7.822

2.2.2.2.1 Standard PCR

The reagents are supplied by KBioscience as PCR buffer, Hot-start Taq and 50mM Mg^{2+}. Primers are obtained separately (Sigma Aldrich, Gillingham, UK). The total reaction volume was 12μL, the final Mg^{2+} concentration 2.2mM as the 4x buffer contains 1.8mM Mg^{2+} (Table 2-9) Reagents were mixed for the total number of samples to be typed and plated onto 96-well optical plates (Applied Biosystems, Woolston, UK), followed by the addition of DNA to each well. Plates were kept on ice when not handled. The plates were sealed with Optical seals and immediately run on a Tetrad PCR machine (MJ Research, Waltham, MA) with a thermal mat to avoid evaporation.

Table 2-10 Conditions for KASPar genotyping

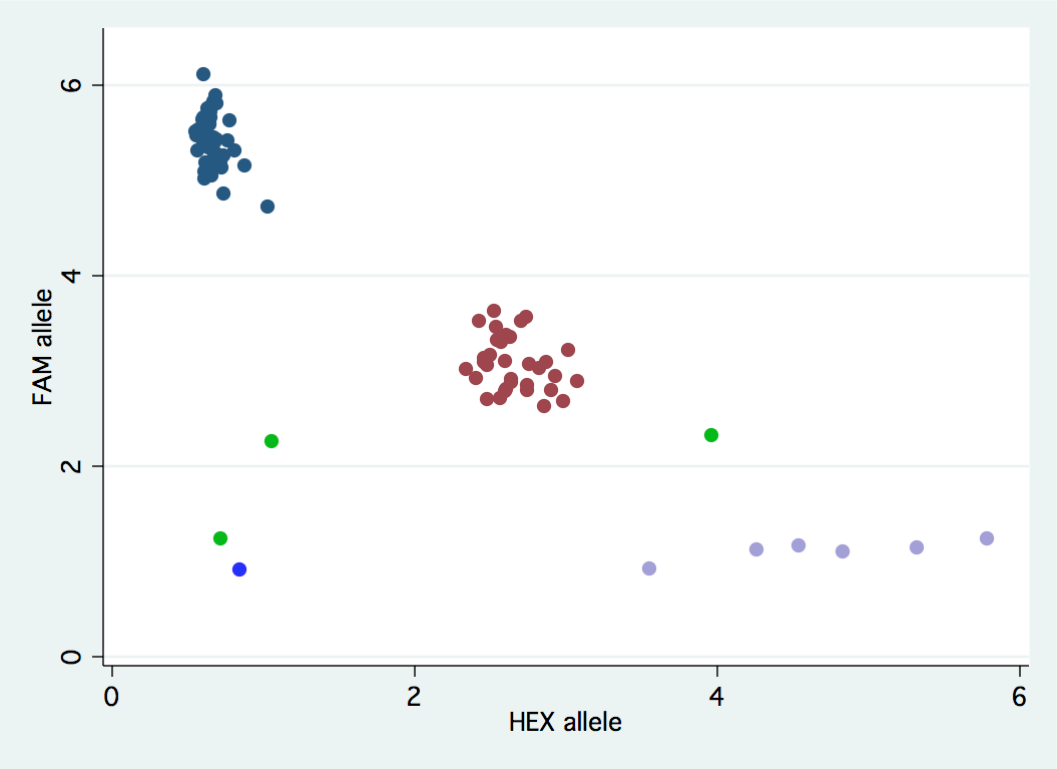
Temperature	Time	
94°C	15 minutes	For Hot-start Taq activation
94°C	10 seconds	
57°C	5 seconds	For 20 cycles
72°C	10 seconds	
94°C	10 seconds	
57°C	20 seconds	For 18 cycles
72°C	40 seconds	
15°C	Until plates removed from machine	

The PCR cycling protocol commenced with a Taq activation step followed by a 2-cycling programme according to manufacturer's instructions. If the incorporation of dye was not fully completed, the second cycling step could be expanded to get better genotyping results (Table 2-10).

2.2.2.2.2 Genotype analysis

Plates were read on an AB7900HT scanner (Applied Biosystems, Woolston, UK), using SDS software 2.2 with automatic genotype calling and 95% confidence intervals. All data were transferred into a single Excel file for cross-checking results between plates and assessing plate heterogeneity by $\chi 2$ test with the appropriate degree of freedom (df=1 for allelic tests, df=2 for genotypic tests). All wells that could not be reliable assigned to a genotype (poor signal, sample lying between clusters) were excluded from further analysis (Figure 2-6).

Figure 2-6 Plot of KASPar dye intensities



Genotype clustering based on FAM and HEX dye intensity. Allele 1 is always the FAM dye, as determined by the PrimerPicker software. The plot shows the results for a 96-well plate, the three genotypes are shown in navy (FAM/FAM), maroon (FAM/HEX), and lavender (HEX/HEX). Blue is the negative control, and green excluded samples that could not reliably be allocated a genotype.

2.2.2.3 LIGHTSCANNER

The LightScanner technology determines genotypes by measuring dye release during DNA melting. The base substitution of a simple SNP subtly alters the melting temperature of the PCR product, which is measured for each well via release of LC

Green incorporated during the PCR. All primers are chosen by the LightScanner® software and are unlabelled.

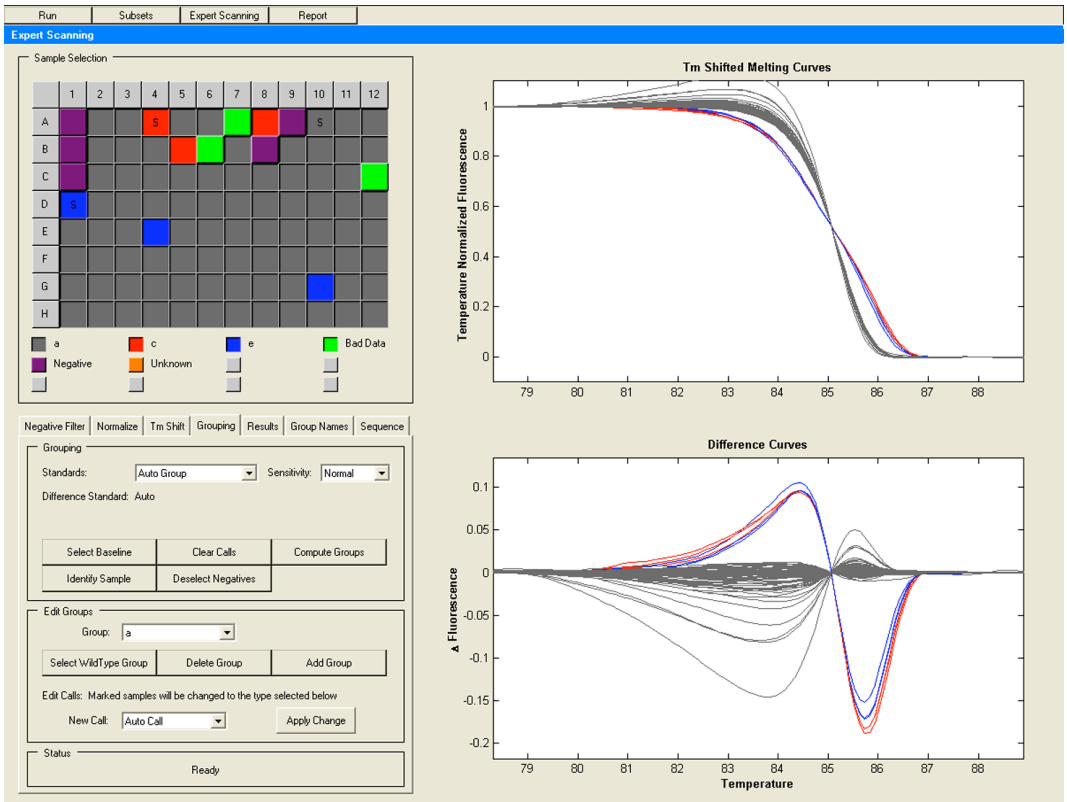
2.2.2.3.1 Standard PCR

The standard PCR was made up to 12.5µl total volume, and typically contained 15-25ng of template DNA per reaction. The Mg2+ concentration was 2.5mM (Table 2-11). Reagents were mixed for the total number of samples to be typed and plated onto white 96-well plates (ABgene, Epsom, UK), followed by the addition of DNA to each well. 20µL of mineral oil were added to each well to avoid evaporation. Plates were sealed with Thermowell sealers (Corning) and immediately run on a Tetrad PCR machine (MJ Research, Waltham, MA). The PCR conditions are given in Table 2-5.

2.2.2.3.2 LightScanner analysis

Following the PCR, plates were loaded into the LightScanner® (Idaho Technologies, Salt Lake City, UT) without seal. Empty wells and controls were marked, and the plate analysed. Groups of the same genotype are indicated by plate position and colour (Figure 2-7).

Figure 2-7 Screenshot of LightScanner analysis software



Bottom left are the controls for analysis. Top left shows distribution of samples in 96-well plate, colour coded by genotype: in this case, grey squares are homozygous wildtype, blue positive controls (mutant allele carriers), red carriers detected in samples. Purple squares are blanks, and green failed samples. The latter two categories are excluded from the curves shown in the right-hand side panels. Top is the temperature normalised fluorescence, the residual fluorescence per well at any given temperature, bottom is the release of fluorescence in each well in response to temperature changes during the LightScanner analysis. The colours are the same as in the left panel, and the difference between wildtype and carriers is evident.

Table 2-11 Reagents for LightScanner PCR

Reagent	Amount [μL]
DNA	1.0
Taq Polymerase	0.125
Forward primer [20mM]	0.125
Reverse primer [20mM]	0.125
MgCl [25mM]	1.25
Buffer	1.25
Nucleotide	1.0
Q solution	2.5
dH_2O	3.875
LC green	1.25

2.3 Statistical methods

All statistical analyses were performed in Stata 9.2 or 10.1 (StataCorp, College Station, TX), p-values were 2-sided and confidence intervals are 95%, χ^2 tests for 2x2 tables were continuity corrected. Fishers exact test was used if expected values in any cell of the χ^2 test were <5. For comparison of a distribution of two variables, the two independent samples t-test was used for normally distributed interval variables (Stata command: ttest), and the Wilcoxon-Mann-Whitney test for all others (Stata command: ranksum). Comparison of medians was performed using a nonparametric K-sample test on the equality of medians (Stata command: median).

2.3.1 Allele naming beyond Illumina

Alleles were represented by A and B according to the Illumina convention (Table 2-3). For A/T and C/G SNPs, the convention was extended that in these circumstances the A allele was A or C, respectively.

2.3.2 Models tested

In all association analyses, the major allele was coded *M* and the minor allele *m*, irrespective of whether the minor allele was A or B. Allele frequencies were determined by

$$M = 2MM + Mm \quad \wedge \quad m = 2mm + Mm \qquad (1)$$

and the allelic OR with 95% CI (the allelic model) was given by

$$OR = \frac{M_{con}}{M_{case}} \times \frac{m_{case}}{m_{con}} \qquad \wedge \qquad SE\log OR = \sqrt{\frac{1}{M_{case}} + \frac{1}{M_{con}} + \frac{1}{m_{case}} + \frac{1}{m_{con}}} \qquad (2)$$

and significance was tested by χ^2 test with df=1.

The genotypic model (genotype frequencies) was tested using a χ^2 test with df=2. In both cases, if the expected frequency in any cell was <5, Fisher's exact test was used. In addition, recessive and dominant models were tested, both are named with respect to the

minor allele *m* compared to the major allele *M*, i.e. the recessive model groups *MM* and *Mm* (vs. *mm*), while the dominant model group *Mm* and *mm* (vs. *MM*). The analyses performed for the latter two models were the same as the genotypic model (Table 2-12).

Genotypes were coded numerically for analysis purposes, *MM*=1, *Mm*=2, *mm*=3. In recessive and dominant models, the group containing the *MM* genotype was coded 0, the other 1. All effect sizes (ES) therefore give the impact of the *m* allele with respect to the *MM* genotype, and dominant and recessive models are with respect to *m* of this naming convention.

2.3.3 Logistic regression

For categorical analyses (case/control, relapse/no relapse), logistic regression was used as the test for trend (Stata command: logistic). Odds ratios generated through logistic regression give a trend estimate for the genotypic model with increasing numbers of *m* alleles (*MM* vs. *Mm* vs. *mm*), while for recessive and dominant models, it is approximated by eq. (2) with the allele counts replaced with group counts. The associated p-value was calculated using Wald's test, where

$$z = \frac{\log OR}{se \log OR} \qquad \wedge \qquad p = 2 \times (1 - normal\,distribution|z|) \qquad (3)$$

2.3.4 Logrank test and Cox regression

In cases where survival data was available, the genotypic, recessive, and dominant models were tested. Time to event analysis was performed using the logrank test for the equality of survivor function (Stata command: sts test). This test formed the basis of the selection of SNPs in Chapter 3 for verification in chapter 4.

For the calculation of the HR for death by Cox regression (Stata command: stcox), alleles were coded in the same fashion as already described (section 2.3.2). All survival data was to first recorded relapse, or right censored for patients lost to follow-up. All follow-up data was left-truncated at the date of surgery. The p_{trend} was calculated using eq. (3) above. The survival data for section 3.4 were initially generated by David Mesher using Stata 10.1.

Table 2-12 Tests of significance used

Genetic model	Components	Comparison of groups	Trend test
Allelic	*M*, *m*	χ^2 test OR based on eq. (1) and (2)	
Genotypic Recessive Dominant	MM vs. Mm vs. mm (*MM+Mm*) vs. *mm* *MM* vs. (*Mm+mm*)	χ^2 test Logrank test	Logistic regression Cox regression

Categorical comparisons were tested using χ^2 or Fisher's exact test, trend testing was performed for the genotypic, recessive, and dominant models, using logistic regression. The Survival functions were tested using the logrank test and Cox regression.

2.3.5 Cochran-Armitage trend test for the additive model

The Cochran-Armitage trend test for the additive genotypic model generates the Υ^2 function, equivalent to χ^2 with df=1, provided λ=1 (see section 2.3.8). Υ^2 is given by

$$\Upsilon^2 = \frac{N\{(r_1 + 2r_2) - \phi(n_1 + 2n_2)\}^2}{\phi(1-\phi)\{N(n_1 + 4n_2) - (n_1 + 2n_2)^2\}} \qquad (4)$$

where N=total number of subjects, ϕ=proportion of cases, r_1=number of *Mm* in cases, r_2=number of *mm* in cases, n_1=number of *Mm* in N, n_2=number of *mm* in N.

2.3.6 Power calculations

Power calculations were performed using the Center for Statistical Genetics, University of Michigan's Power Calculator for Genetic Studies (CaTS). This is designed for power calculations for 2-stage genetic association studies based on the principles described by Skol *et al.*[274], see Table 10-1 for the URL.

2.3.7 Determination of Hardy-Weinberg equilibrium (HWE)

Deviation from HWE was tested for by genotypic χ^2 test with df=1, calculating the expected frequencies from the observed allele frequencies, and based on

$$a^2 + 2ab + b^2 = 1 \qquad (5)$$

where a=allele frequency of A, and b=allele frequency of B.

HWE was tested for the overall group, control group, and case group, with $p<0.01$ deemed to be a significant deviation in the overall and control groups. Cases were allowed to deviate from HWE, as this could be manifestation of the association of the tested locus with the case group.

2.3.8 Detection of population substructure

Deviation from the expected versus the observed quantiles of the χ^2distribution were analysed visually using Q-Q plots. The genomic inflation factor λ was given by

$$\lambda = \frac{median(\chi^2_{obs})}{median(\chi^2_{\exp})} \qquad (7)$$

with 0.456 the predicted median χ^2 for df=1, and 1.386 for df=2, and the reference line adjusted accordingly[275].

The Eigensoft programme was used to test for ancestral variation between cases (relapsed patients) and controls (non-relapsed patients). Eigensoft is a software suite to correct for population stratification using a three-step algorithm: first, principal component analysis (PCA) is used to infer genetic variation along principal axes based on ancestral differences between cases and controls, and the output displayed graphically and in list format[276]. Secondly, if significant stratification is detected, then genotypes are adjusted to take

account of ancestral differences and thirdly, the new genotypes are used for future association studies[277].

The PCA step is very sensitive to known population stratifiers, e.g. HLA types. To avoid these known stratifiers obscuring the potential signal from non-random distribution within the VICTOR cohort, a random set of SNPs was chosen, spaced evenly across the genome. This method reduces the impact of the HLA types by using fewer SNPs residing in this area, without decreasing the power such that cryptic relatedness between subjects could not still be detected. This analysis was performed by Dr Jean-Baptiste Cazier.

2.3.9 Linkage equilibrium testing

Linkage disequilibrium relationships were determined using Haploview 4.1 (ref [278]). LD blocks were determined using the algorithm described by Gabriel *et al.*[279], as implemented in Haploview 4.1. For all risk and outcome loci, r^2 was used as the determinant of the LD relationship. For the section on tagging insertion/deletion changes (Chapter 8), *D'* was used to capture the LD relationship of rare alleles more accurately.

r^2 is given as the square of Pearson's correlation coefficient, r is given by

$$r(X,Y) = \frac{Cov(X,Y)}{SD(X) \times SD(Y)} \qquad (8)$$

D' is calculated using allele frequencies at both loci and is a ratio of observed LD (denoted *D*) and maximal LD (denoted D_{max}) between the two loci[280]. Assuming that the allele frequencies of the four alleles (A, a and B, b) are given by p_1, p_2, q_1, and q_2, then the gamete frequencies g_{11}, g_{10}, g_{01}, and g_{00} are given as

			Locus A	
	Alleles		A	a
		Allele frequency	p_1	q_1
Locus B	B	p_2	g_{11}	g_{01}
	b	q_2	g_{10}	g_{00}

Then *D* is given by

$$D = g_{11}g_{00} - g_{10}g_{01} \qquad (9)$$

and $D=0$ if there is complete linkage equilibrium. *D'* is given by

$$D' = \frac{D}{D_{\max}} \qquad (10)$$

where

$$D_{\max} = \min(p_1q_2, q_1p_2) \text{ for } D \geq 0 \quad \wedge \quad D_{\max} = \min(p_1p_2, q_1q_2) \text{ for } D < 0 \qquad (11)$$

As such, *D'* is also high when the minor allele homozygote of one locus is (almost) always associated with the minor allele homozygote of the second locus even if the reverse was

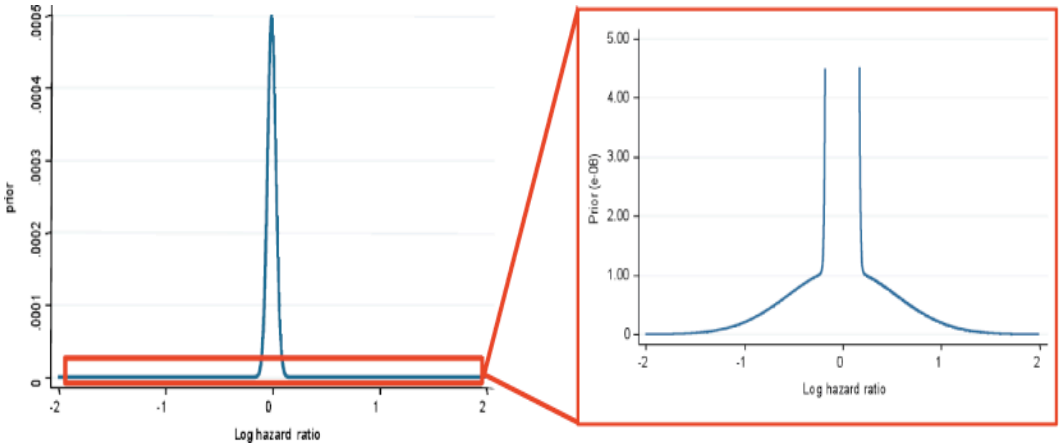
not true. In these instances r^2 would be lower, suggesting no LD. In practice, this requires inference of haplotype frequencies in a given dataset, a step inherent in the analysis performed by Haploview 4.1.

2.3.10 Shrinkage of estimates and their standard errors

The first step was to make assumptions about the true underlying associations between the SNPs and relapse from CRC: that at most hundreds of SNPs would have a true hazard ratio of greater than 1.15 (or less than 0.87, the reciprocal of 1.15), that at most 50 would have a true hazards ratio of greater than 1.2 (or less than 0.83), and that most of the true HR would be close to 1.0.

Formally, therefore, for the logHR, all but 3 per 10,000 would be normally distributed with mean zero and standard deviation 0.04. In addition, a small proportion (1.5 per 10,000) was to have a moderate positive association and the same small proportion to have a moderate negative association. Formally, both would also be normally distributed with means +/- 0.2 and standard errors of 0.5. The prior distribution on the log hazard ratios was the mixture of these three normal distributions. With 309,200 SNPs, this prior probability gave 113 hazard ratios greater than 1.15, 36 greater than 1.2, 22 greater than 1.5 and 9 greater than 2. By symmetry, the same numbers were less than the reciprocal of these hazard ratios.

Figure 2-8 Graphical Representation of prior probability density function



The prior probability function assumes that virtually all SNPs have no effect on prognosis (logHR≈0), and that only a tiny proportion have logHR≠0, representing SNPs with an effect on prognosis.

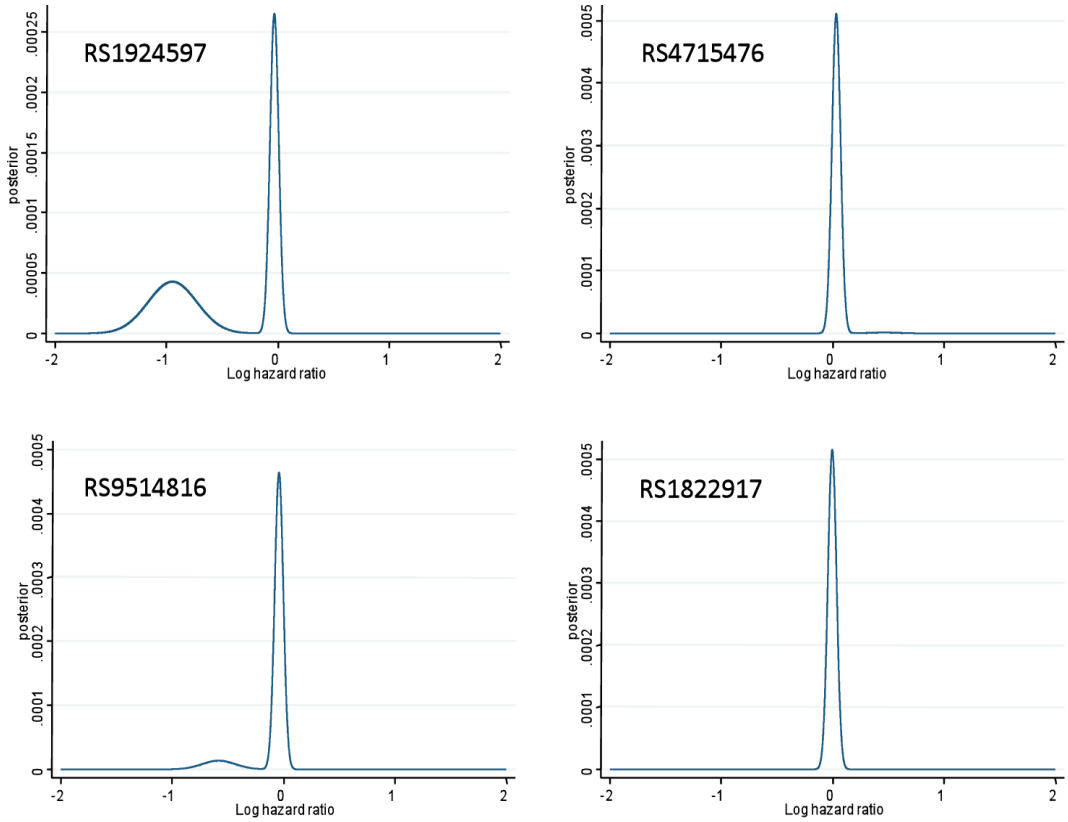
The prior probability density function is illustrated in Figure 1. It looked like a spike at zero, but if zoomed in and ignoring the function between -0.175 and 0.175, it was apparent that the tails had "shoulders" at about +/- 0.2, corresponding to the possibility of containing a few clinical important associations (Figure 2-8).

The clinical data yielded estimated log hazard ratios and standard errors for each SNP. These were treated as representative of the means and standard errors of normally distributed likelihoods.

These were combined with the prior distribution to obtain a posterior distribution for each SNP (Figure 2-9 for examples). Due to the unusual prior chosen - i.e. the combination of three normally distributed SNP populations rather than assuming that all SNPs lie within the same normal distribution - the posterior density function can be bimodal, with two spikes separated by a region with very little probability mass. Despite this, the posterior means and standard errors and used in the meta analysis were treated as if they came from normally distributed posteriors, and the main spike adjusted for the magnitude and probability of the second peak.

The shrinkage was performed by David Mesher and Prof Peter Sasieni, and applied to the analyses in section 4.3.

Figure 2-9 Examples of posterior probability density functions



Examples of the posterior probability for the dominant model of four SNPs. Both examples on the left have an adjusted HR<1 (logHR<0), with logHR for rs1924597 being -0.466, and for rs9514816 being -0.095. The greater deviation from zero for rs1924597 is driven by the greater posterior probability of logHR=-1. The SNPs on the right show a logHR≈0 (0.038 for rs4715476 and -0.003 for rs1822917), with no evidence of a second peak.

2.3.11 Meta-analysis

All meta-analyses of OR or HR were performed using log transformed effect sizes in a fixed effects model with inverse variance weighting. Meta-analyses of counts (e.g. of alleles) were performed using a fixed effects model with Mantel-Haenszel weighting.

Heterogeneity was assessed using the I^2 statistic, namely the proportion of observed variation in effect size attributable to heterogeneity (Q) given by

$$Q = \sum_i \frac{1}{\mathrm{var}(\log HR_i)} (\log HR_i - \log HR)^2 \qquad \wedge \qquad I^2 = 100\% \times \frac{Q - (k-1)}{Q} \qquad (6)$$

where $\log HR_i$ is the log transformed point estimate of the individual study, logHR the summary estimate, and $\mathrm{var}(\log HR_i)$ the variance of the point estimate for the individual study.

Values of $I^2 \geq 50\%$ were considered moderate heterogeneity[281]), and the use of a random-effects model was considered. Likewise if qualitative measures of study characteristics revealed inter-study heterogeneity. Heterogeneity was further assessed with Egger's bias coefficient[282] and by funnel plot[283].

When performed, sensitivity analysis to exclude a significant influence of other characteristics was performed by meta-regression (Empirical Bayes model)[284].

Chapter 3

GWAS for prognostic markers: Discovery phase

The patients of the VICTOR trial offered a unique opportunity to employ a GWAS approach to biomarker discovery: to date, there has been very limited study of the germline in the prognostication and prediction of response in CRC, and the VICTOR trial combines the availability of good quality DNA with robust clinical data. This GWAS was conceived in parallel with the successful GWAS for CRC risk loci performed in our laboratory, which has already demonstrated the successful application of high throughput, unbiased SNP screening for a given phenotype[42-46].

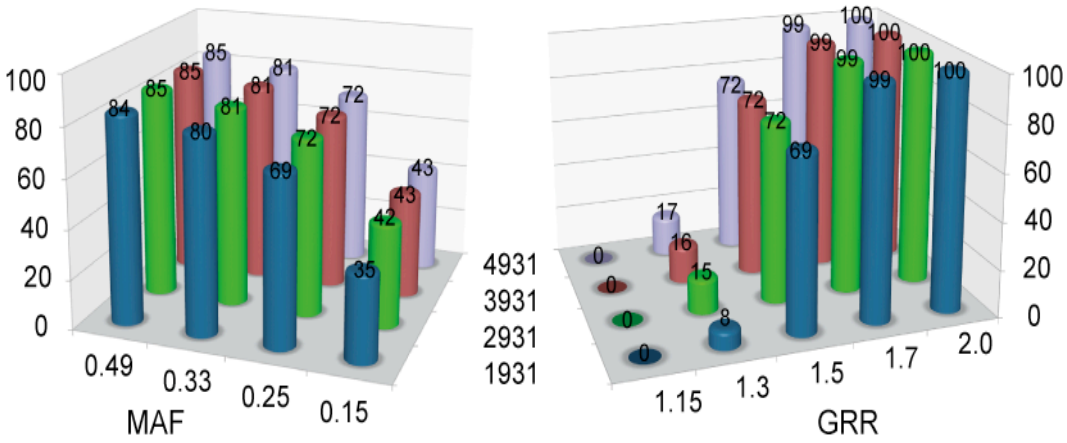
3.1 Study design

The VICTOR GWAS for novel markers of outcome in CRC was designed as a 2-stage study: an initial phase (discovery phase or phase 1) during which all autosomal SNPs (309,200) on Illumina Hap300 arrays would be tested for an association with outcome, and a verification phase (phase 2), for the top 40 SNPs of interest. The results from both phases were planned to be combined to derive a 'final' estimate of prognostic information, and to confirm or exclude the selected SNPs as prognostic markers. The incorporation of the hypothesis generating set in the verification analysis improves power despite more stringent p-values for statistical significance[274]. At the beginning of the discovery phase, a further cohort of stage 2 and 3 patients from a clinical trial (PETACC 3) had already been identified, and further cohorts would be considered as they became available.

The chosen strategy provides a trade-off between statistical power and cost[274], and while it would be preferable to type all SNPs in all samples of the verification phase as this is the power limiting step, the cost associated with this approach is prohibitive. Assuming a second cohort of 1000 patients, an overall event rate of 25% (phase 1 and 2 combined), a

frequency of the allele associated with relapse of 33% in the overall population, this study would have 85% power to detect a relative risk of 1.5 at a cut-off of p≤1e-07 in the joint

Figure 3-1 Power estimation for different genotyping scenarios



Power estimation based on the cohort sizes, and various assumptions for relative risk and allele frequency. Both models assume an overall relapse rate in all patients of 25%. The cohort size is given in the central column, with 931 patients in phase 1, and between 1000 and 4000 additional patients in phase 2. The model on the left assumes a constant genotypic relative risk (GRR) of 1.5 and varying MAF. The model on the right assumes constant MAF of 0.25 and varying GRR. Power estimations are based on the Power Calculator for Genetic Studies[274].

analysis. With the same assumptions, but a second cohort of 2000 patients, the power increases only marginally, as is the case with a further 3000 and 4000 additional patients. With an allele frequency of 25%, only with relative risks above 1.5 does power approach 80%, almost irrespective of the number of additional patients (Figure 3-1). Therefore, this study was powered to detect either quite marked effects (relative risk ≥1.5), or risk alleles with quite high frequencies (≥30%).

3.2 Patient details

At the cut-off point in November 2007, 934 samples of the VICTOR trial were available, with 156 relapses (cases) and 778 non-relapsed patients (controls). There was a slight preponderance of males; the median age of diagnosis was 64.4 years; these variables were well matched between stages. There was a slightly higher proportion of rectal carcinomas in the stage 3 group. Therapy was as expected for site and stage of disease, with significantly higher rates of adjuvant chemotherapy and relapse in stage 3 (Table 3-1).

In three cases the last date known to be alive was not available in the database under this heading; these patients were allocated a censor data of the last known contact with the trials office, the date of discontinuation of the trial medication. Two patients were allocated to the 15th day of the months as relapse date, as this part of the date was missing. Fourteen patients died of other causes and were censored. One patient died on relapse; this was considered the relapse date.

Table 3-1 Patient characteristics of analysed patients

	All	Stage 2	Stage 3	p-value
Patients	931	475	459	
Gender [%male]	64.7	65.7	63.6	5.11e-01
Median age at diagnosis (range)	64.4 (24.6-86.4)	64.9 (24.6-86.4)	64.2 (34.0-83.9)	1.47e-01
Disease site				1.00e-02
Colon [%]	64.8	68.2	61.2	
Rectosigmoid junction [%]	8.3	9.1	7.4	
Rectum [%]	27.0	22.7	31.4	
Adjuvant chemotherapy [%]	62.6	31.1	95.0	4.96e-90
Radiotherapy [rectal patients %]	38.7	35.5	41.0	3.80e-01

p-value is for the comparison between stage 2 and 3, for interval variables the two independent samples t-test was used, and the Wilcoxon-Mann-Whitney test for all others

3.3 Parameters and quality control of genotyping

3.3.1 Genotype quality based on internal controls

In total, 934 patients were analysed using the Illumina Infinium Duo arrays: 526 using Hap300 arrays, and 408 using Hap370 arrays. The parameters used to assess the quality of the data were GenTrain score for each individual SNP, call rates for each sample, and GenCall score for individual genotypes per patient and SNP (see section 2.2.1.2 for a detailed description).

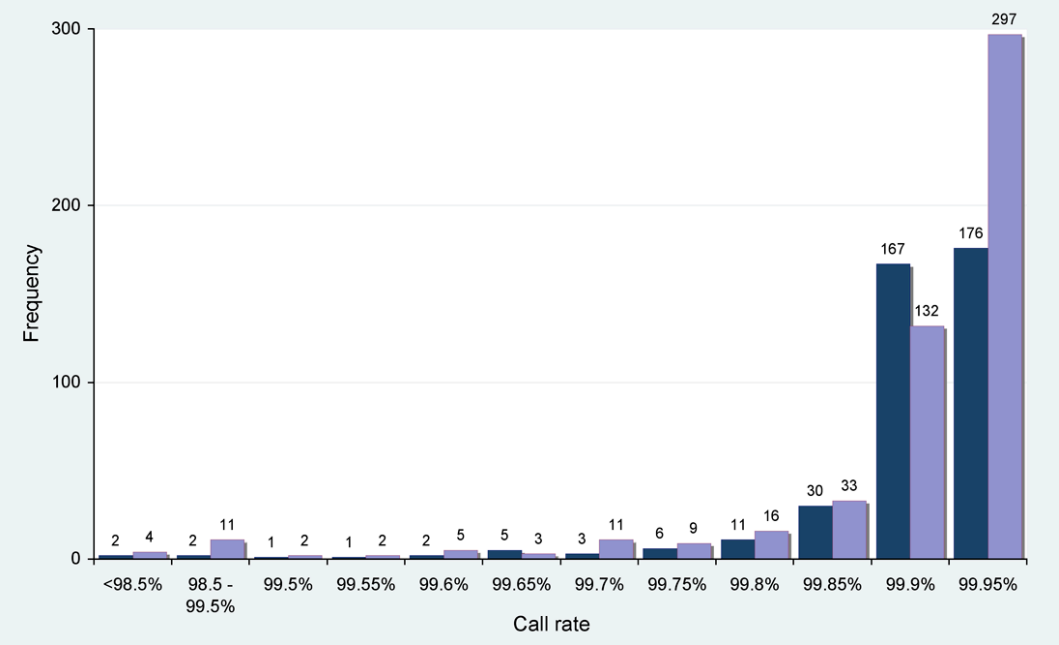
3.3.1.1 PER SAMPLE ANALYSIS

Samples were only included in the analysis if the call rate was >95%, this was the case for 931 samples. Three samples had call rates <95%: in one case, this was due to the sample quality, as a repeat of this sample had the same poor call rate (65%), and in two cases this was likely to be a failed SNP array, as both samples were analysed on the same array and had equally poor call rates (70%). There was insufficient DNA to repeat these samples. One patient was enrolled twice in the trial (as VIC0284 and VIC0890) and only the higher scoring array was included in the analysis.

For the patients included, the median call rate was 99.93% (range 96.65-99.99%) and the mean call rate was 99.90%, with only six samples having a call rate <98.5%, and 19 samples <99.5%; the majority of samples had a call rate>99.90% (Figure 3-2). The GenCall score was not assessed at a per genotype level at this stage, but was taken into account when the SNPs for further investigation were selected.

The proportion of call-rates ≥99.9% was higher for Hap370 than Hap300 arrays, while the proportion of samples with a call-rate ≥99.95% was higher in the Hap300 arrays, 56.6% vs. 43.4% (Table 3-2). From a practical point of view, however, call rates of ≥99.90% for each array are good enough to accept the array as not failed and the genotypes generated as valid.

Figure 3-2 Distribution of patients by call rate



The percentages below the bars are the lower bounds of 0.05% bins for the call rate, the exact value is included in that bin. The bottom two bins are for call-rates 95 to <98.5% and 98.5 to <99.5%. All values are absolute numbers for Hap300 (light blue) and Hap370 (navy) of the samples included in the analysis.

3.3.1.2 PER SNP ANALYSIS

Single nucleotide polymorphisms were only considered for further analysis if >95% of samples had a genotype allocated by the genotyping software (Illumina BeadStudio 3.0). If this was less, the SNP was deemed to have failed (98 SNPs). For the remaining SNPs, the median number of samples called per SNP was 931(range 885-931), and in 97.82% at least 926 samples were called. Genotyping per SNP was robust: the median GenTrain score for the VICTOR samples was 0.848 (range 0.367-0.972). In total, 159 SNPs, equivalent to 0.05% of all SNPs, were excluded and had their GenTrain score set to zero.

Table 3-2 Call rate by array type

	All	Hap300	Hap370	p-value
Median call rate	99.95%	99.95%	99.95%	0.492
Call rate ≥99.90%>	82.92%	81.71%	84.48%	2.01E-13
Call rate ≥99.95%	50.81%	56.57%	43.35%	4.75E-12

The call rate distribution differs significantly by array type (Hap300 vs. Hap 370) without having any known practical implications.

While the GenCall score was not analysed for individual SNPs and samples at this stage, the summary GenCall score for each SNP (based on the distribution of the individual GenCall scores) suggested that the samples performed uniformly well and genotyping was very robust. When the median GenCall score (GC_{50}) for each SNP was compared to the 10th centile GenCall score (GC_{10}) for the same SNP, the values were identical in all but

774 SNPs, suggesting that, at a minimum, the best performing 90% of samples for each SNP behave in a uniform fashion during genotype calling.

3.3.2 Repeat samples and reproducibility

Out of the 931 samples, 17 had call rates at initial genotyping of just below the 95% threshold required for inclusion in further analysis, they all repeated with call rates ≥99.90% on 2nd genotyping. The reasons for failing to achieve the required threshold initially were mostly due to technical problems with the first batch of Hap300 arrays, when occasionally bubbles would be introduced while loading the DNA onto the array. Later Hap300 arrays and the Hap370 arrays did not suffer from this problem.

To test the reported robustness of the Infinium chemistry employed in the Hap 300 and Hap370 arrays[270], data for the one duplicate sample with excellent call-rates in both runs (VIC0890 and VIC0284), and one SNP (rs6983267), also typed by direct sequencing, were analysed. VIC0284 was typed once on the Hap300 and once on the Hap370 arrays (as VIC0890 in the latter). For 555 SNPs (0.17%), the genotypes were not concordant, the majority (475 SNPs, 85.6% of non-concordant SNPs) because the SNP was a 'no-call' in one of the samples. Of the remaining SNPs, all but one were AB calls for VIC0284 (Hap300) and BB calls for VIC0871 (Hap370). The one exception was a the reverse, BB in VIC0284, AB in VIC0871, suggesting a slight difference in the calling of genotypes for SNPs where the AB and BB clusters are close together. Overall, however, for the Hap300 content, the concordance was >99.97%, with only 80 SNPs allocated a different genotype in the two runs. None of the SNPs later taken into verification (Chapter 4) fell into this category.

Table 3-3 Discrepancies between direct sequencing and Illumina genotyping

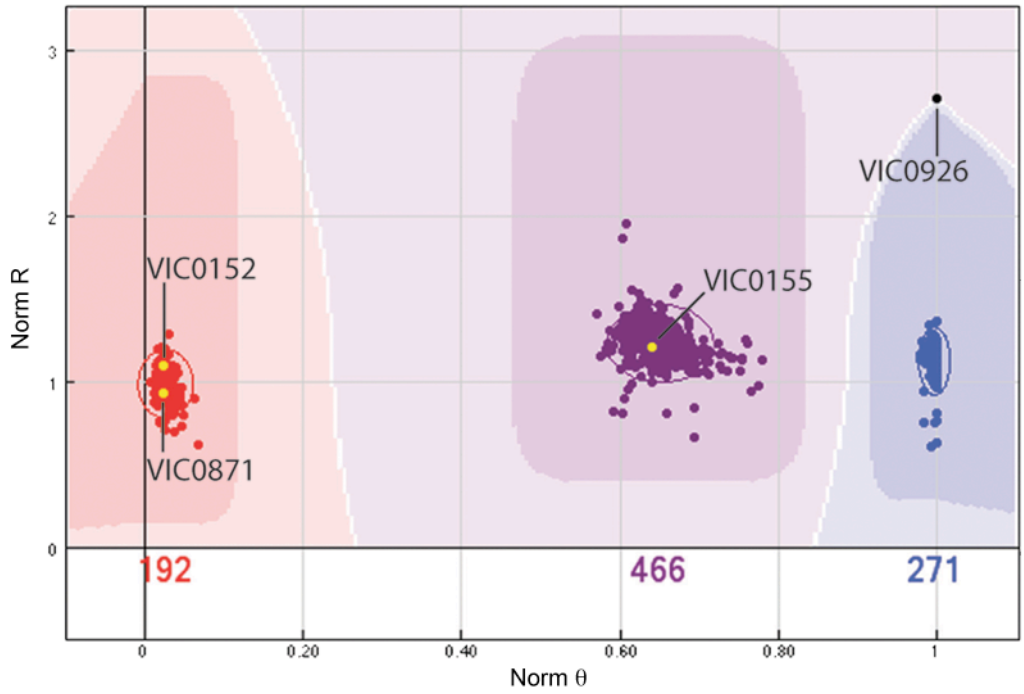
Sample ID	Illumina	GenCall score	Sequencing
VIC0152	TT	0.874	TG
VIC0155	TG	0.874	GG
VIC0871	TT	0.874	TG

Genotypes for rs6983267 in samples where direct sequencing differed from the Illumina called genotypes. In all cases, the Illumina genotypes appeared robust on genotyping parameters (GenTrain score=0.842 and GenCall scores=0.874 for all three samples).

rs6983267 was typed in 881 samples for which data was available from the Illumina arrays as part of the initial GWAS for CRC risk[44]. Manual genotyping of the same sample had been performed by Dr Oliver Sieber using direct sequencing. Genotypes were called blinded to the results from the other technology, and reproduced in 878 samples (99.7%). In three cases, the genotype was discrepant by one allele (Table 3-3). The genotypes for all three samples when called by BeadStudio are robust, they all have the same high GenCall score (0.874) and there is no ambiguity about the cluster membership (and thus

genotype) based on the 15 replicates per array (Figure 3-3). It is likely that the discrepancy arose on the side of direct sequencing, and Illumina genotypes were used in all cases.

Figure 3-3 Genotype clusters for rs6983267



Genotype clusters for rs6983267 with discordant samples highlighted in yellow. Based on the 15-fold replication inherent in the Illumina arrays, there appears no doubt about the genotypes for these samples, making it likely that the discrepancy arose on the side of the sequencing. Sample VIC0926 was excluded from the survival analysis even though on visual inspection it looks most likely to be a BB genotype.

3.3.3 Manual recalling of genotypes

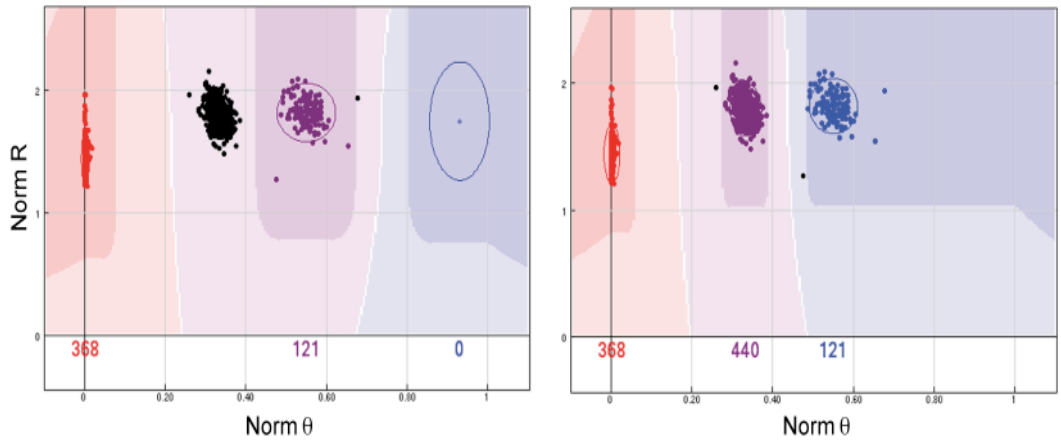
While the Illumina genotyping algorithms attempt to " . . . *mimic evaluations made by a human expert's visual and cognitive systems*."[272], Illumina recommend manual adjustment of SNP clusters where distinct clusters form but the algorithm has been unable to determine these with certainty.

All SNPs with a call frequency with of <98.71% were inspected and adjusted manually where samples were not allocated a genotype despite clearly belonging to a particular cluster. The cut-off was driven by the Illumina standard of expecting a successful SNP to have a call frequency of >98.5%, and because the yield for successful increases in call frequency for SNPs with a call frequency between 98.5% and 98.71% had already been low. Therefore, the next step up in call frequency (98.82%) was not assessed.

The aim of manual recalling was to maximise the call frequency, without calling any unreliable genotypes, i.e. clustering had to be unequivocal for the adjusted samples. This was performed for 1855 SNPs and improved the call frequency for the majority of SNPs inspected. In a small minority of SNPs, manual recalling decreased the call frequency as

the cluster separation had not been satisfactory on visual inspection and equivocal samples had their genotype calls removed, or SNPs were deemed to have failed (Figure 2-4, left).

Figure 3-4 Example of manual calling of clusters



The genotypes for rs10904494 clearly fall into distinct clusters, but the calling software does not recognise the rightmost cluster as the BB cluster as it is not close enough to the expected position (Norm θ=1, left). Manual recalling assigns appropriate genotypes to all three clusters. This manoeuvre increases the call frequency from 52.5% to 99.8% for this SNP (right).

The impact of manual recalling on global parameters of genotyping, however, was modest, as only 0.6% of SNPs were inspected in this fashion (1855 of 309,200 SNPs) , and for those edited, for many, the call frequency improved by only a few samples per SNP, although for some SNPs, the call frequency improved much more dramatically (Figure 3-4).

3.4 Disease-free survival analysis

Only autosomal SNPs (309,200) were analysed for an association with survival, based on the 931 samples and the follow-up available at the data cut-off point on 17 November 2008. All SNPs were analysed for the four models: genotypic, dominant, recessive, and allelic. They were analysed by logrank test for equality of the survival function and Cox regression with a time-to-event analysis for the genotypic, dominant and recessive models, and by allelic OR for the allelic model.

Table 3-4 Survival and relapse

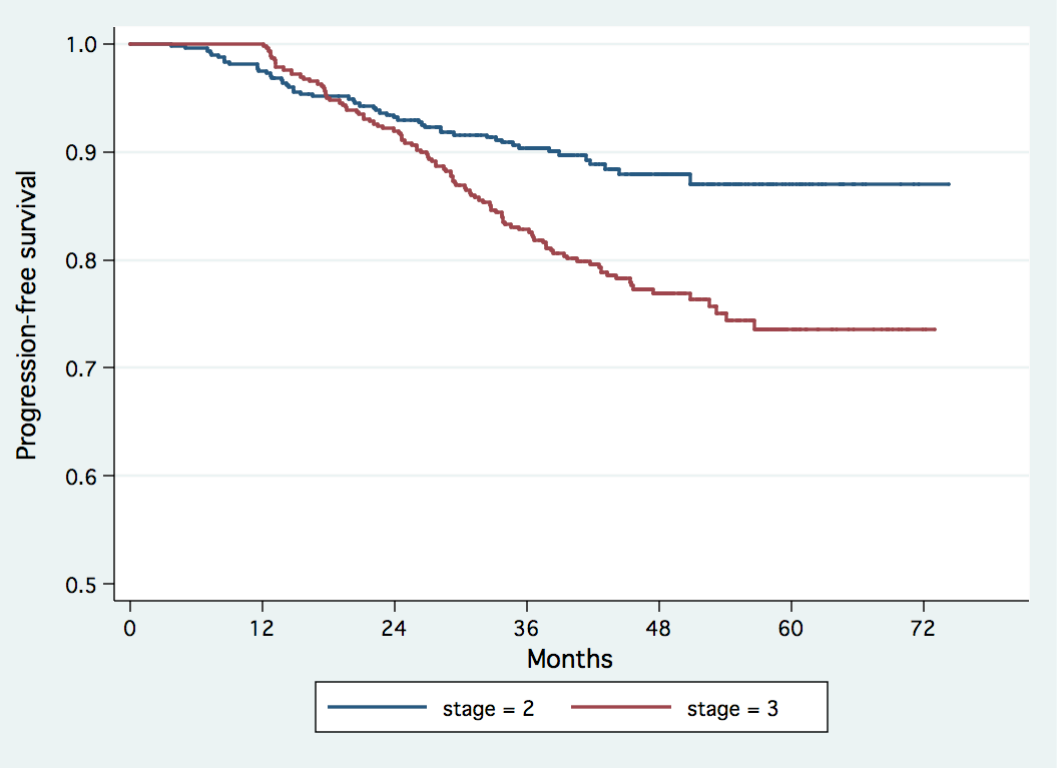
	All	Stage 2	Stage 3	p-value
3-year DFS [%]	86.5	90.4	82.8	2.07e-03
All patients	931	472	459	
All relapsed patients in follow-up	156	51	105	
All relapsed patients in follow-up [%]	16.8	10.8	22.9	8.50e-05

3-year DFS appears higher than would be expected from the relapse rate due to a further 34 relapses beyond 36 months follow-up. The p-value is for Cox regression comparing stage 2 and 3.

All survival data were to first recorded relapse, or right censored for patients lost to follow-up. The median follow-up was 44.1 months, SD ± 12.8 months for the overall cohort, and 46.1 months, SD ± 10.3 months if relapsed patients are excluded.

Table 3-4 shows the number of relapsed patients during the whole follow-up period and gives the estimated 3-year DFS, by stage. 156 patients had relapsed (16.8%), and as expected, this was higher in stage 3 than stage 2 (22.9% and 10.8%, respectively; p=8.50e-05). The median survival was not reached at the data cut-off, with a 3-year disease-free survival of 86.5%. For stage 2, 3-year DFS was 90.4% and for stage 3, DFS=82.8%, with a HR=1.79 (95% CI 1.23-2.59, p=2.07e-03). The 3-year DFS does not take into account all recorded relapses in the VICTOR cohort, and there were a further 34 relapses beyond 36 months follow-up (Figure 3-5).

Figure 3-5 Survival by stage in VICTOR cohort

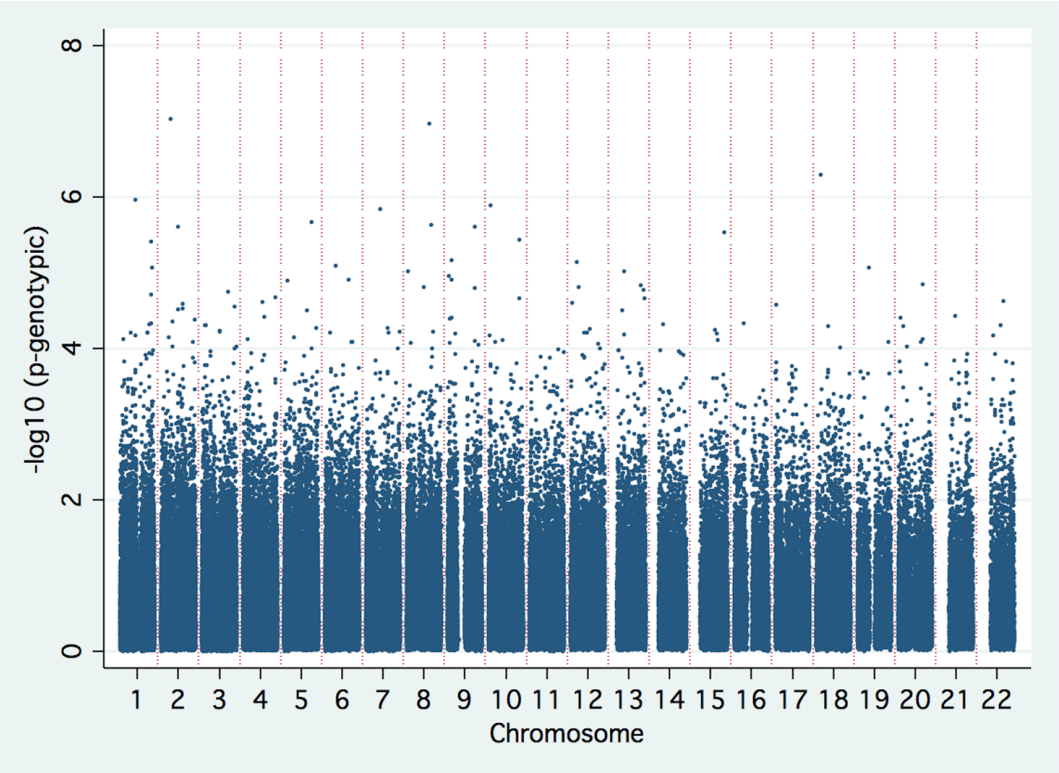


As would be expected, there is a statistically significant difference in the survival between the stages (p=8.50e-05); stage 2 - navy; stage 3 - maroon.

3.4.1 Genome wide p-values

The distribution across the genome of p-values for the SNPs analysed was assessed visually by plotting the negative logarithm of the p-value from the logrank test for the genotypic model against SNP position on each chromosome (Manhattan plot). Figure 3-6 shows the plot for the whole chromosome.

Figure 3-6 Genome wide Manhattan plot



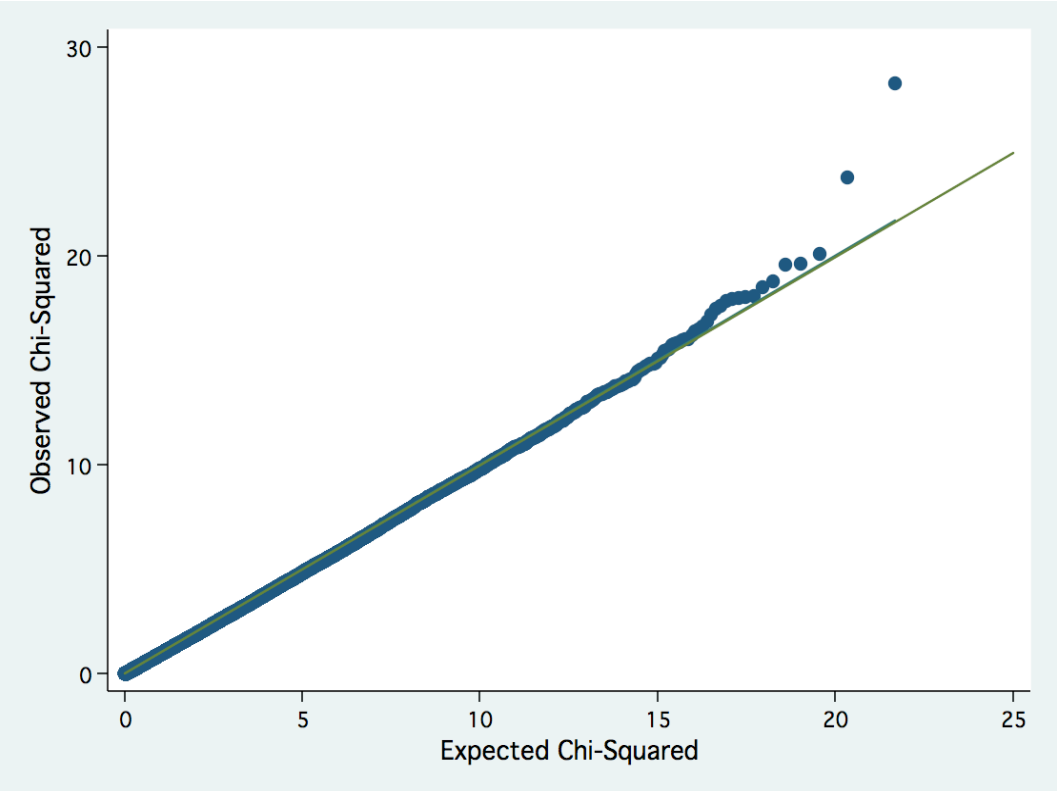
Plot of the negative $\log_{10}$ of p_{gen} vs. the position on chromosomes 1 to 22, starting at the telomeric end of the p-arm of chromosome 1 and ending at the telomeric end of the q-arm of chromosome 22. There are small gaps to indicate the different chromosomes, where there are larger gaps, there are no SNPs in that region (the short arm of chromosomes 13, 14, 15, 21 and 22; the centromeric region of chromosomes 9, 16, and 18).

The Manhattan plot for each chromosome was scrutinised for regions of nominal significance ($-\log_{10}(p_{gen})>4.0$). Areas where there were adjacent SNPs with a $-\log_{10}(p_{gen})>3.0$ were also noted, likely to represent areas of LD with a possible association with prognosis . This process was to ensure that the selection strategy based on numeric values (outlined in section 3.6) did not miss potentially significant regions. All Manhattan plots by chromosome are given in Appendix B.

3.5 Population stratification

Hidden population substructures are capable of artificially inflating the $\chi 2$ statistic for positive associations, causing spurious assertions about the link between a particular marker and outcome[285,286]. To test if the observed matched the expected $\chi 2$ distribution, the inflation factor λ was calculated and the distribution shown in a Q-Q-plot of observed versus expected quantiles of the $\chi 2$ distribution for the allelic model (Figure 3-7). On visual inspection, there was no evidence of systematic overestimation of the $\chi 2$ statistic. The inflation factor was $\lambda_{all}=1.001$, again suggesting no hidden substructure. There was no evidence in the other models (genotypic, recessive, dominant) for an inflation of the $\chi 2$ statistic ($\lambda_{gen}=0.989$, $\lambda_{rec}=1.010$, $\lambda_{dom}=0.999$).

Figure 3-7 Q-Q plot for allelic model



Q-Q plot of observed against the expected quantiles of the χ^2 distribution. The green line represents the perfect fit of data and model, in other words, exact agreement between the two distributions. The deviation from the perfect fit line at the tail is expected, as there is a lower density of points in this part of the plot, and 'true' associations would be expected to move the observed distribution above the expected.

The genetic substructure was further analysed using principal components analysis (PCA) as part of the Eigenstrat software suite designed to detect population substructure. PCA aims to reduce a large number of measurements, e.g. genotypes, to a few principal components which can explain the majority of the variability, an approach largely applied to the detection of sub-populations in admixture mapping, but which can be equally applied to any population and trait[276]. "*Thus, the first principal component is the mathematical combination of measurements that accounts for the largest amount of variability in the data.*" [287].

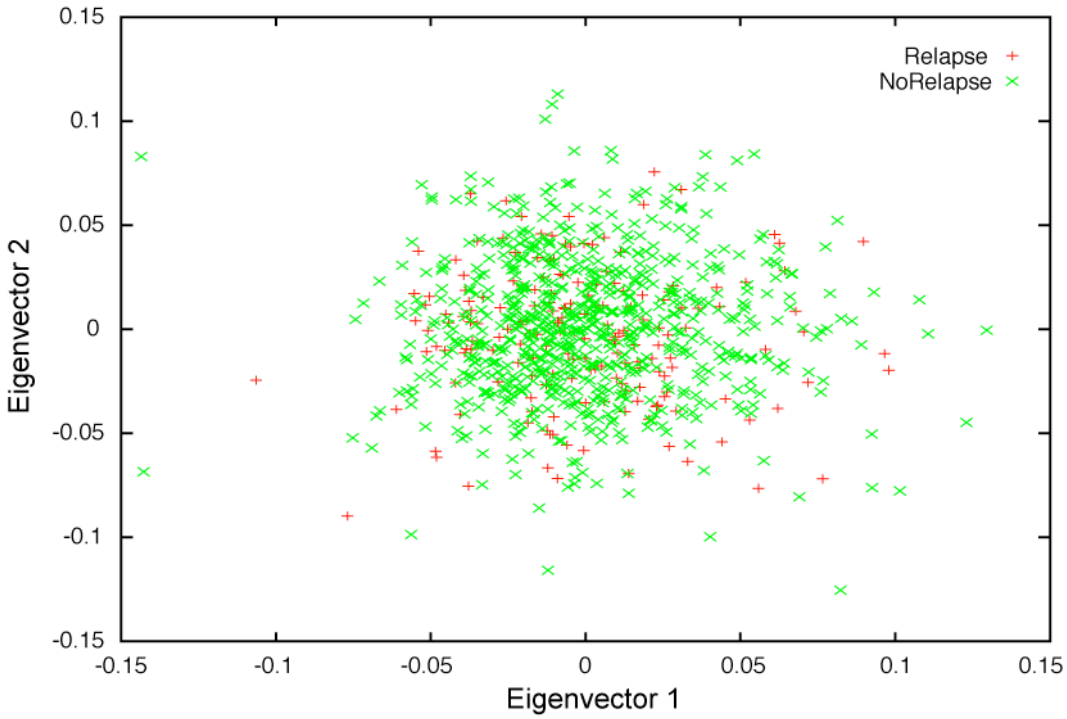
Table 3-5 p-values associated with the top 10 Eigenvectors

Eigenvector	p-value	Eigenvector	p-value
1	0.570	6	0.741
2	0.127	7	0.362
3	0.857	8	0.782
4	0.297	9	0.356
5	0.560	10	0.568

The p-values for each Eigenvector are non-significant, meaning that there is no significant separation along the vector.

PCA was performed by Dr Jean Baptiste Cazier, and did not reveal any population stratification along the top ten Eigenvectors as assessed by numerical means (Table 3-5) and visual inspection (Figure 3-8). The p-values associated with individual SNPs (see section 3.4 for the survival analysis and p-value generation) were therefore not adjusted prior to selection of SNPs to be taken into the verification phase.

Figure 3-8 Eigenvectors 1 and 2 for VICTOR cohort



Plot generated by SmartPCA, the PCA component of the Eigenstrat software suite[277]. Eigenvectors are named 1, 2, 3, . . . N, in decreasing order of impact on variability, and vectors 1 and 2 are shown. There is an even distribution of the samples around the mean for both vectors, implying that there is no heterogeneity of the population. Lower order Eigenvectors should in this instance also have even distributions, and visual inspection confirmed this (see Appendix B).

3.6 SNP selection for verification

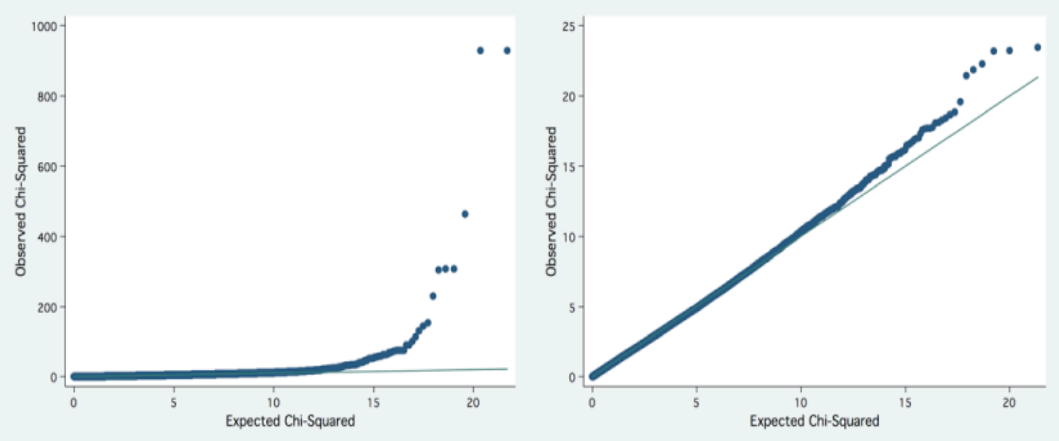
3.6.1 Selection strategy and rationale

Forty SNPs were to be taken into the initial verification phase. These were selected from the autosomal content of the Hap300 arrays after excluding failed and non-informative SNPs, those not in HWE, and those not meeting the significance threshold for inclusion of p<1e-04.

The logrank test was more sensitive than $\chi 2$ and Fisher's exact test to small numbers in the minor homozygote group mm, with marked inflation of the observed logrank $\chi 2$ values. Therefore, SNPs with low minor homozygotes frequencies were excluded, reducing the over-inflation of positive associations. The higher the cut-off value for the minor homozygote frequency, the lower the deviation of the observed tail on the Q-Q

plot, but this risks excluding true associations. The cut-off chosen was a minor homozygotes frequency less than 10, and while the appearance of the Q-Q plot was markedly improved (Figure 3-9), only in the genotypic model did it improve λ appreciably (new λ_{gen}=0.996).

Figure 3-9 Omission of SNPs with low minor homozygote frequencies



The χ^2 distribution derived from the logrank test appeared to be more sensitive to small numbers in the minor homozygote group with marked inflation of the observed χ^2 values (left). Therefore, SNPs with minor homozygotes less than 10 in the control group were not considered further, producing a less inflated distribution (right). Model shown is recessive.

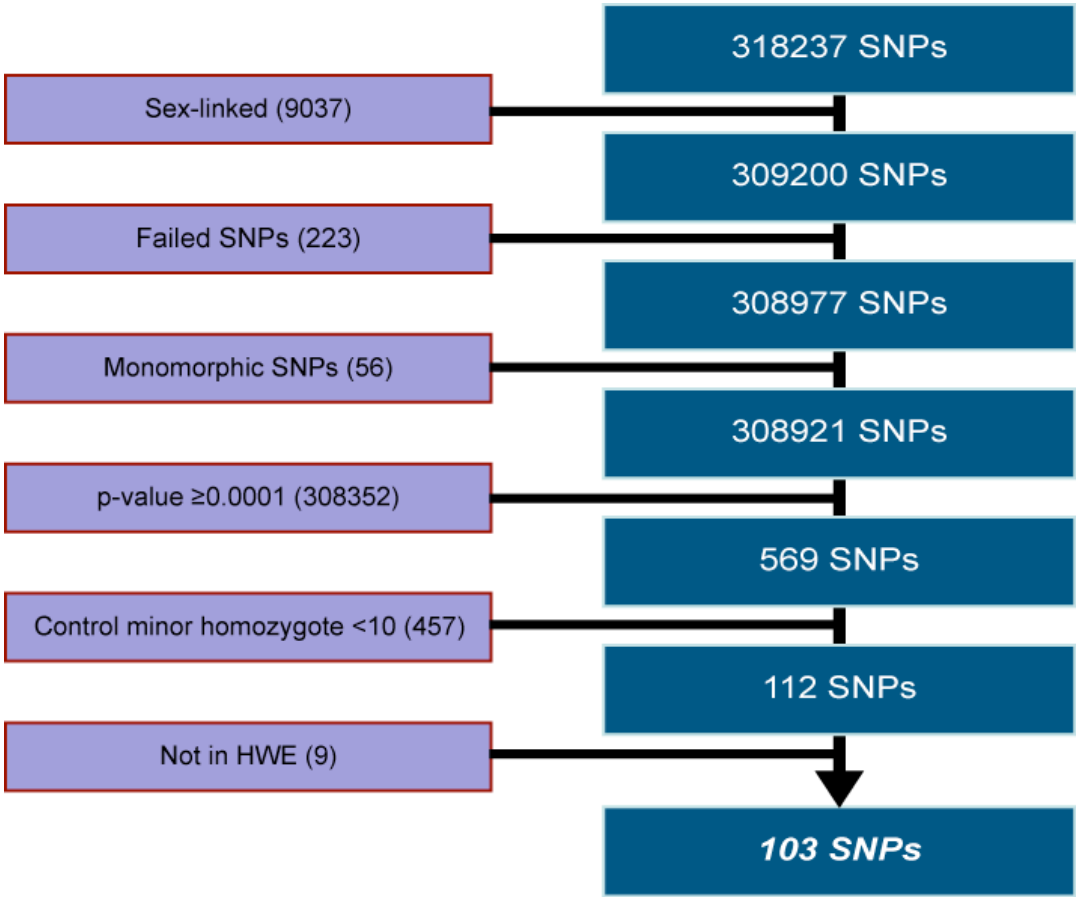
Twenty five SNPs would be chosen purely based on statistical significance; 15 SNPs would be chosen from the remainder of the eligible SNPs based on biological rationale.

Biological rationale was assessed using the function of the gene nearest to the SNP as denoted by the UCSC Genome browser , e.g. involved in cell signalling, cycling, and motility, have a previous described role in or association with cancer, the distance of the gene to the SNP, level of conservation in region of the SNP, some evidence for other SNPs in the region being associated with prognosis based on Manhattan plots, and finally, if two SNPs appeared to have a similar biological rationale, statistical relevance (p-value). Further, in order to maximise the number of distinct loci to be evaluated in phase 2, if two SNPs were in the same gene, only the SNP with the lower p-value would be taken forward, provided they did not have markedly differing genotype distributions, in other words low linkage disequilibrium (LD) defined as r^2<0.7. The reasons for choosing individual SNPs based on biological rationale are described in section 3.6.2.

Thus, of the 318,237 SNPs on the arrays, 9037 were not included in the survival estimation as they were X or Y chromosome linked, leaving 309,200 autosomal SNPs with survival analysis data. 180 SNPs were excluded because they had a GenTrain score of 0, and a further 43 SNPs because they did not cluster well enough for any further analysis. These 223 SNPs constituted the failed SNPs. 56 SNPs were excluded because they were monomorphic in the VICTOR cohort. After applying the p-value cut-off, 565

SNPs remained, excluding 308,356 SNPs. 457 further SNPs were excluded because a minor homozygote frequency of <10, leaving 112 SNPs with a MAF of >0.12. Nine SNPs were not in HWE in the control group or overall cohort, leaving 103 SNPs. All performed well on genotyping parameters, with a median GenTrain score of 0.843 (range 0.708-0.937) and a median call frequency of 931 (range (927-931). For the complete list, see Table 10-3.

Figure 3-10 Flowchart of SNP selection



Starting with 318,237 SNPs, these were reduced in a succession of steps according to the pre-planned exclusion strategy, numbers excluded in brackets. The final 103 SNPs formed the basis of selection for the verification phase.

The genotypes and quality parameters for the chosen SNPs are given in Table 3-6, and the distinction between SNPs chosen statistical significance and biological rationale is indicated. The impact on survival and a functional annotation is given in Table 3-7 (see also Table 10-2).

Table 3-6 Genotypes for SNPs selected for verification in phase 2

		Control				Relapse						HWE		
Rank	SNP	AA	AB	BB	Total	AA	AB	BB	Total	GT	GC50	p_{total}	$p_{control}$	Alleles
1	rs472660	9	169	597	775	12	39	105	156	0.853	0.887	0.234	0.440	C/T
2	rs7556894	38	291	445	774	24	55	77	156	0.861	0.898	0.649	0.274	A/G
3	rs2589183	46	252	476	774	2	26	128	156	0.901	0.937	0.034	0.106	C/T
4	rs764372	11	199	565	775	11	41	104	156	0.784	0.791	0.931	0.162	A/G
5	rs672757	105	386	284	775	42	83	31	156	0.848	0.881	0.206	0.146	A/G
6	rs7912136	386	329	59	774	77	49	30	156	0.853	0.888	0.356	0.332	A/G
7	rs6972789	218	399	158	775	36	61	59	156	0.830	0.859	0.754	0.318	C/T
8	rs7007146	217	391	167	775	31	65	60	156	0.866	0.903	0.544	0.712	C/T
9	rs745888	44	320	411	775	25	60	71	156	0.838	0.870	0.618	0.072	A/G
10	rs4978394	575	188	12	775	97	47	12	156	0.927	0.958	0.526	0.447	A/G
11	rs3784780	14	223	537	774	11	23	122	156	0.760	0.751	0.723	0.093	C/T
12	rs10510044	410	303	62	775	61	64	31	156	0.839	0.871	0.088	0.567	C/T
13	rs712082	27	228	518	773	18	41	96	155	0.781	0.785	0.030	0.751	A/C
14	rs11788150	78	329	367	774	35	61	60	156	0.804	0.821	0.104	0.735	A/C
15	rs10842099	377	323	75	775	104	47	5	156	0.842	0.874	0.462	0.632	C/T
16	rs4715476	527	228	19	774	85	59	12	156	0.823	0.849	0.708	0.330	A/G
17	rs1526884	389	314	72	775	108	36	11	155	0.889	0.926	0.062	0.455	A/G
18	rs4649314	81	370	322	773	35	60	59	154	0.791	0.801	0.754	0.093	A/C
19	rs1350308	119	355	300	774	45	72	39	156	0.851	0.886	0.143	0.410	A/C
20	rs9533457	413	318	44	775	78	54	24	156	0.853	0.887	0.830	0.088	A/G
21	rs1571583	29	297	449	775	19	61	76	156	0.880	0.917	0.191	0.018	C/T
22	rs567564	62	324	388	774	30	63	63	156	0.918	0.951	0.502	0.621	A/G
23	rs7866165	30	291	454	775	18	51	87	156	0.856	0.891	0.522	0.046	G/T
24	rs437171	18	240	517	775	15	48	93	156	0.832	0.862	0.890	0.107	A/G
25	rs3752261	523	237	15	775	93	51	12	156	0.807	0.826	0.335	0.044	A/G

Table 3-6 Genotypes for SNPs selected for verification in phase 2 (cont'd)

		Control				Relapse						HWE		
Rank	SNP	AA	AB	BB	Total	AA	AB	BB	Total	GT	GC50	p_{total}	$p_{control}$	Alleles
26	rs1924597	565	193	17	775	139	15	2	156	0.886	0.923	0.432	0.914	A/G
27	rs1822917	447	292	36	775	90	44	21	155	0.936	0.964	0.647	0.178	C/T
28	rs9514816	10	155	610	775	6	53	97	156	0.843	0.875	0.876	0.965	A/C
29	rs6518956	284	345	144	773	44	100	12	156	0.810	0.831	0.808	0.032	A/G
30	rs1878632	269	383	123	775	51	59	46	156	0.810	0.830	0.449	0.491	G/T
31	rs7600624	590	171	14	775	96	50	10	156	0.932	0.961	0.226	0.694	A/G
32	rs9315425	13	202	560	775	10	31	115	156	0.931	0.960	0.589	0.281	C/T
33	rs10429965	599	166	10	775	119	28	9	156	0.868	0.905	0.172	0.693	T/C
34	rs5997921	11	164	599	774	9	35	112	156	0.863	0.899	0.174	0.953	G/T
35	rs1824476	70	336	369	775	5	51	100	156	0.852	0.886	0.697	0.602	C/T
36	rs4776494	574	188	12	774	95	53	8	156	0.862	0.898	0.754	0.444	C/T
37	rs1438620	259	361	154	774	27	85	44	156	0.833	0.862	0.326	0.167	A/G
38	rs2514841	172	403	200	775	13	94	48	155	0.895	0.932	0.025	0.250	C/T
39	rs1034116	14	168	592	774	8	51	97	156	0.807	0.826	0.358	0.605	C/T
40	rs25689	26	266	483	775	14	67	75	156	0.816	0.839	0.271	0.145	A/G

Genotypes and associated data for 40 SNP selected for phase 2. SNP chosen on statistical grounds only (see text) are highlighted in grey. The GenTrain score is for that SNP, the SNP_{50} score is the 50^{th} centile of the GenCall scores for that SNP.

Table 3-7 Survival and associated genes for SNP selected for phase 2

Rank	SNP	Chr	Position	Nearest gene	Distance	Logrank p	Model	ES (95% CI)	p-trend
1	rs472660	7	99298043	CYP3A43 protein	intronic	1.62e-09	recessive	5.11 (2.83-9.22)	6.21e-08
2	rs7556894	2	65365572	Actin-related protein 2 isoform b	15kb	9.19e-08	recessive	3.09 (2.00-4.77)	3.98e-07
3	rs2589183	8	97591685	Syndecan 2 precursor	intronic	1.06e-07	allelic	0.37 (0.25-0.55)	1.00e-06
4	rs764372	18	9817368	RAB31, member RAS family	intronic	5.09e-07	recessive	4.24 (2.30-7.84)	4.00e-06
5	rs672757	1	112148779	K+- voltage-gated channel	intronic	1.08e-06	allelic	1.85 (1.45-2.38)	4.53e-07
6	rs7912136	10	6790287	Protein kinase C, theta	128kb	1.27e-06	recessive	2.59 (1.74-3.85)	3.04e-06
7	rs6972789	7	66757993	Stromal antigen 3-like 4	350kb	1.42e-06	recessive	2.18 (1.57-3.01)	2.57e-06
8	rs7007146	8	103603973	E3 ubiquitin protein ligase	108kb	2.33e-06	recessive	2.14 (1.55-2.95)	4.01e-06
9	rs745888	2	119626784	Complement component 1	4kb	2.48e-06	recessive	2.69 (1.75-4.12)	6.09e-06
10	rs4978394	9	111306488	PTPN3 protein	6kb	2.48e-06	genotypic	1.72 (1.32-2.24)	5.08e-05
11	rs3784780	15	89110479	Bloom's helicase	intronic	2.94e-06	genotypic	0.88 (0.64-1.20)	4.17e-01
12	rs10510044	10	120227066	Receptor for prolactin-releasing hormone	130kb	3.60e-06	recessive	2.46 (1.66-3.64)	7.45e-06
13	rs712082	1	222792683	EST BU947851	intronic	3.89e-06	recessive	3.01 (1.84-4.92)	1.13e-05
14	rs11788150	9	13815565	Nuclear factor I/B	250kb	6.77e-06	recessive	2.32 (1.59-3.38)	1.24e-05
15	rs10842099	12	23080053	Homo sapiens cDNA FLJ37414 fis	intronic	7.26e-06	allelic	0.51 (0.37-0.69)	1.55e-05
16	rs4715476	6	54730244	Tubulointerstitial nephritis antigen	370kb	8.12e-06	genotypic	1.73 (1.33-2.24)	3.81e-05
17	rs1526884	19	22095491	near cluster of ZNF proteins	31 - 750kb	8.47e-06	dominant	0.47 (0.33-0.66)	1.38e-05
18	rs4649314	1	231636367	Mixed lineage kinase 4	50kb	8.54e-06	recessive	2.30 (1.57-3.35)	1.52e-05
19	rs1350308	8	4600956	CSMD1: CUB and Sushi multiple domains 1	intronic	9.39e-06	allelic	1.75 (1.37-2.22)	1.19e-05
20	rs9533457	13	42806849	Ecto-NOX disulfide-thiol exchanger 1	intronic	9.62e-06	recessive	2.58 (1.67-3.99)	2.01e-05
21	rs1571583	9	4257209	GLIS family zinc finger 3 isoform b	intronic	1.10e-05	recessive	2.80 (1.73-4.53)	2.58e-05
22	rs567564	6	116263486	FRK tyrosine kinase	105kb	1.21e-05	recessive	2.37 (1.59-3.53)	2.20e-05
23	rs7866165	9	13802685	Nuclear factor I/B	273kb	1.23e-05	recessive	2.85 (1.75-4.67)	2.94e-05
24	rs437171	5	18165562	Cadherin 18, type 2 preproprotein	1.3Mb	1.28e-05	recessive	3.08 (1.81-5.25)	3.45e-05
25	rs3752261	20	44553077	Zinc finger protein 334 isoform b	10kb	1.41e-05	recessive	3.41 (1.89-6.15)	4.48e-05

Table 3-7 Survival and associated genes for SNP selected for phase 2 (cont'd)

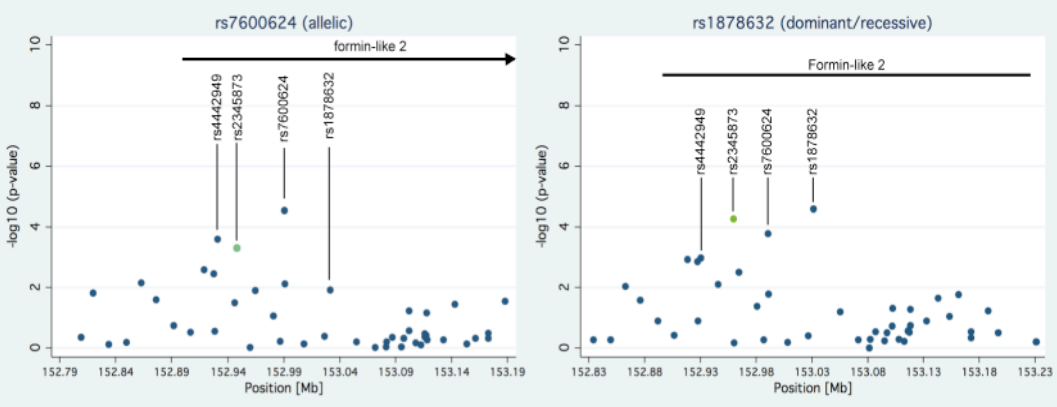
Rank	SNP	Chr	Position	Nearest gene	Distance	Logrank p	Model	ES (95% CI)	p-trend
26	rs1924597	13	96238572	Heparan sulfate 6-O-sulfotransferase 3	intronic	1.46e-05	allelic	0.38 (0.23-0.61)	5.81e-05
27	rs1822917	8	72430567	Eyes absent 1 isoform c, protein tyrosine kinase	intronic	1.55e-05	recessive	2.65 (1.67-4.21)	3.26e-05
28	rs9514816	13	107581123	DNA ligase IV	77kb	1.68e-05	allelic	2.08 (1.52-2.86)	4.69e-06
29	rs6518956	22	34271966	RASD family, member 2	intronic	2.37e-05	genotypic	0.97 (0.78-1.21)	7.95e-01
30	rs1878632	2	153031404	Formin-like 2	intronic	2.58e-05	recessive	2.06 (1.46-2.91)	3.84e-05
31	rs7600624	2	152990621	Formin-like 2	intronic	2.91e-05	allelic	1.96 (1.45-2.66)	1.60e-05
32	rs9315425	13	35980608	Cyclin A1	65kb	3.09e-05	recessive	3.58 (1.89-6.81)	9.74e-05
33	rs10429965	1	209907121	NEK2 protein kinase	exonic	4.78e-05	recessive	3.68 (1.88-7.22)	1.50e-04
34	rs5997921	22	29960896	LIM domain kinase 2 isoform 1	intronic	4.88e-05	recessive	3.68 (1.88-7.22)	1.52e-04
35	rs1824476	18	27281399	Desmoglein 3 preproprotein	0.6kb	5.10e-05	allelic	0.55 (0.41-0.74)	6.47e-05
36	rs4776494	15	62497880	Thyroid hormone receptor interactor 4	exonic	5.66e-05	genotypic	1.77 (1.35-2.32)	3.55e-05
37	rs1438620	2	222746705	PAX3	26kb	8.25e-05	dominant	2.25 (1.49-3.41)	1.28e-04
38	rs2514841	8	109129843	R-spondin family, member 2	intronic	9.83e-05	recessive	0.34 (0.19-0.60)	2.05e-04
39	rs1034116	14	89024929	Checkpoint suppressor 1	intronic	1.14e-04	allelic	1.89 (1.39-2.56)	6.19e-05
40	rs25689	9	73549916	Transmembrane protein 2	exonic	1.81e-04	allelic	1.69 (1.30-2.22)	9.40e-05

Survival function of SNP selected for verification phase. The p-value is by logrank test on which the SNP was chosen or Fisher's exact test for the allelic model, and the effect size (ES) and 95% CI are from Cox regression for the same genetic model (OR in the case of allelic model). For SNP chosen based on the allelic model, the significance of the trend test, p_{trend} from the Cochran-Armitage test. The nearest gene is either the one in which the SNP is located (exonic or intronic), or the nearest gene in either direction on the chromosome. Unless the SNP is located within a hypothetical gene, the nearest gene is always a known gene (see also sections 3.6.2 and Table 10-2).

3.6.2 Choosing on biological rationale

Following the top 25 most significant SNPs, rs1822917 and rs9514816 were the next most significant, excluding rs6487867 because this was not in HWE in the control group in the initial calculation, although is in fact only borderline (p_{HWE}=0.011), and rs1411684 because it was in high LD with rs4978394 already included. Of the following next eight most significant SNPs, only one SNP (rs6518956) was included as it was located in a gene (RASD family, member 2, part of the family of RAS related proteins), while a further SNP (rs4411217) of these eight was in high LD (r^2=0.72) with a SNP already included (rs10510044).

Figure 3-11 Manhattan plot for rs1878632 and rs7600624



Two SNPs located in formin-like 2 were included in the verification phase (rs1878632 and rs7600624), they were most significant in different models: rs1878632 in the recessive model (top) and rs7600624 in the allelic model (bottom). The plots revealed that there was one other SNPs in formin-like 2 that was of borderline significance (rs4442949), and a further SNP excluded based on low homozygote frequency but with a p-value <1e-05 is indicated in green.

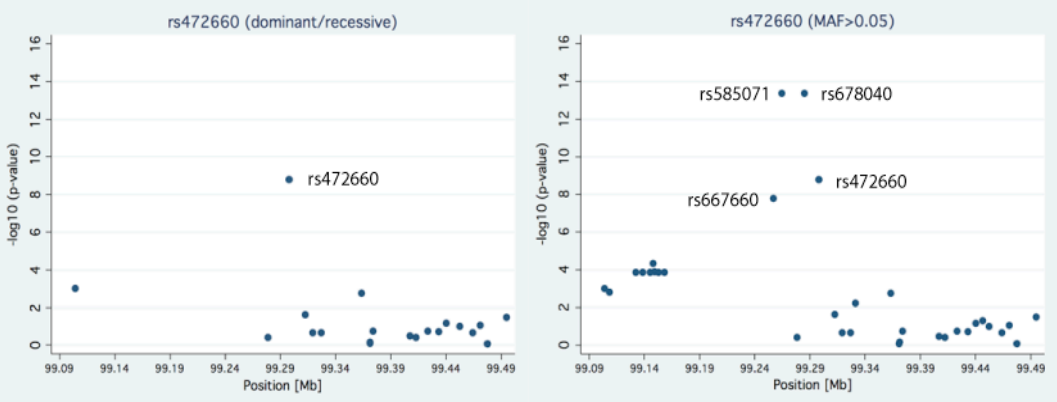
The next included SNPs are rs1878632 and rs7600624, both located in formin-like 2 on chromosome 2, whose overexpression has been linked to CRC metastasis[288]. The two SNPs were of similar significance even if the effect on survival was more marked for rs7600624 (HR_{gen}=1.77, 95% CI 1.36-2.29, p_{gen}=1.80e-05). They are only in weak LD (r^2=0.22) and not in the same LD block, although located in the same intron. There was one other SNP (rs4442949) in formin-like 2 that was of borderline significance with p_{all}=2.56e-04, and an excluded SNP based on low homozygote frequency (rs2345873) with a p_{rec}=3.12e-05 (Figure 3-11).

rs9315425 was included because it is in the vicinity of cyclin A1, a known regulator of the cell cycle. A further four SNPs were chosen because they are located within or very close to genes: rs10429965 is a non-synonymous change in *NEK2*, a protein kinase that is abnormally expressed in many cancers and may have a role in mitotic regulation and chromosomal instability[289]; rs5997921 is intronic in LIM domain kinase 2, a gene implicated in tumour invasion and metastasis formation[290]; rs1824476 immediately

upstream of desmoglein 3 and further upstream, the desmoglein cluster, members of the cadherin superfamily of cell adhesion molecules, implicated as markers of progression[291] and worse prognosis in oral squamous cell cancer[292]; and rs4776494 in *TRIP4*, implicated as a transactivator of NF-κB[293]. Based on the regional Manhattan plots, three (rs10429965, rs5997921, rs1824476) of these five SNPs had some evidence of other SNPs in the vicinity with an association with prognosis as supporting evidence of possible biological relevance.

rs1438620 is located in *PAX3*, a known transcription factor associated with the development of alveolar rhabdomyosarcoma[294] and the development of borderline and malignant phyllodes tumours of the breast[295]. rs2514841 is located in *RSPO2*, an r-spondin associated with *WNT* in the canonical signalling pathway[296]. Both these SNPs have Manhattan plots supporting other SNPs in the region as being associated with prognosis.

Figure 3-12 Manhattan plot for rs472660



The Manhattan plot for rs472660 on chromosome 7 does not suggest any other SNPs associated with outcome in the region when the rare variant cut-off is set to minor homozygote frequency ≥9 (left). This is equivalent to setting the MAF=0.12 If the rare variant cut-off is set to MAF>0.05, several other SNPs suggest that this region is highly significantly associated with prognosis (right) and three further highly significant SNPs emerge (rs585071, rs678040, and rs667660).

Four SNPs did not reach the required level of significance on logrank testing, but had a p_{trend} <1e-04. Two, rs1034116 and rs25689, were chosen as they were located within genes (checkpoint suppressor 1 and transmembrane protein 2, respectively).

Lastly, rs472660, located intronic in *CYP3A43*, was one of the six SNPs with a minor homozygote count of 9, excluded on the basis of a low homozygote count, despite missing the cut-off by only one. It was included as it was the most significant association of these six SNPs (and any of the 40 chosen), and a cytochrome P450 detoxification gene would be an interesting candidate for a prognostic marker. While the Manhattan plot based on the SNPs as selected did not suggest other SNPs in the region to support an association of rs472660 with outcome, setting the cut-off for rare variants to MAF>0.05 would have

included many highly significant SNPs, supporting the inclusion of rs472660 in the verification set (Figure 3-12). Lastly, it lies within a strong LD block covering *TRIM4* and *GJE1* (gap junction protein epsilon 1), another candidate from the cell adhesion functional family, while *CYP3A4* upstream of rs472660 is not covered by the same LD.

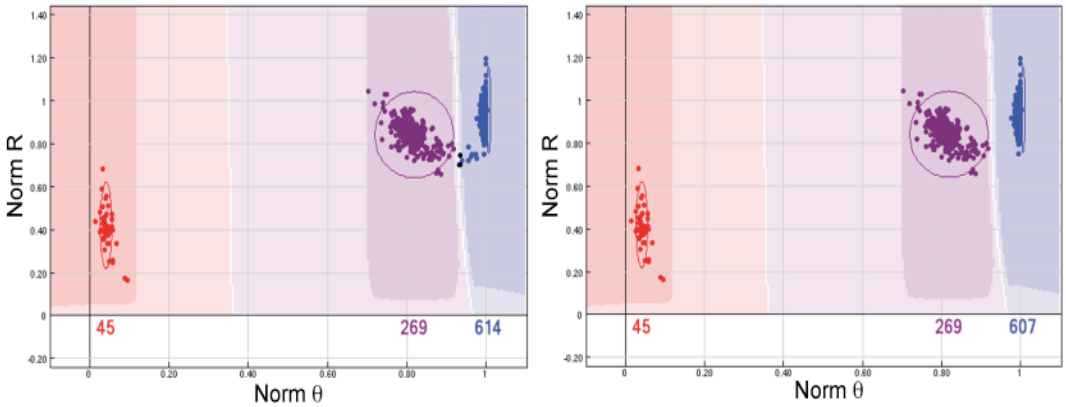
3.6.3 Summary of SNPs taken forward into Phase 2

The selected 40 SNPs represented a well performing subset of the 103 SNPs, based on genotyping parameters: the clustering was tight, with a median GenTrain score of 0.852 (range 0.76-0.936). Consequently, the number of samples called per SNP was high, with at least 927 samples per SNP (median 931, SD ±0.871) called. This is equivalent to a call frequency per SNP ≥99.6%.

The median GenCall score for all called genotypes of the 40 SNPs taken into phase 2 was 0.886 (range 0.157-0.964), with only 50 genotypes out of 37,240 genotypes for 40 SNPs and 931 samples having a GenCall score of <0.7.

As expected, all 40 SNPs were in HWE for the overall and control cohorts at the 0.01 level. For the relapsed cohort, 12 SNPs (rs472660, rs7912136, rs3784780, rs712082, rs1526884, rs9533457, rs1822917, rs6518956, rs1878632, rs9315425, rs10429965, rs2514841) were not in HWE, with rs3784780 being the most significant deviation (Table 3-6).

Figure 3-13 Clustering of rs712082 before and after adjustment



Three samples do not have a genotype call assigned by the calling software (left), but another seven are deemed unreliable on visual inspection and are excluded (right).

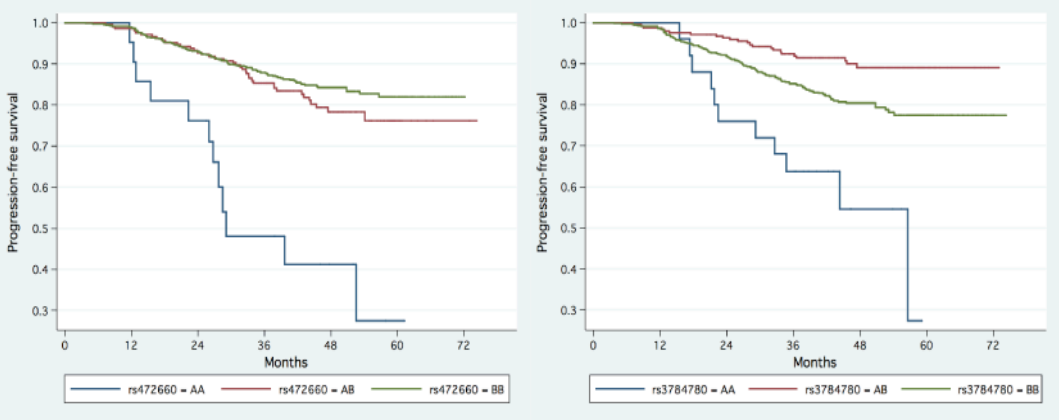
All genotype plots were inspected visually for unreliable clustering. rs712082 was the only SNP where on visual inspection, cluster separation was not optimal between AB and BB clusters (Figure 3-13). As a consequence, rs712082 had the 2nd lowest GenTrain score of any of the selected SNPs. Three samples were excluded by the calling software, while seven were called but were deemed unreliable on inspection of the genotype plot. When the unreliable genotypes were excluded from the survival analysis, the resulting HR and

95% CI were essentially unchanged but slightly more significant (HR_{rec}=3.03 (95% CI 1.85-4.96, p_{rec}=9.87e-06) compared to HR_{rec}=3.01 (95% CI 1.84-4.92, p_{rec}=1.13e-05) with all samples. Excluding the samples from the HWE calculations did not alter the outcome, with both overall and control cohort displaying no significant deviation from HWE. For further analysis, the unreliable samples remained excluded.

There were no issues with the cluster definition in any of the other SNPs, including the SNP with the lowest GenTrain score (rs3784780). In that case, while the clusters were not as tight, there was no ambiguity about the cluster membership of individual samples.

There were four SNPs for which the p_{trend} was greater than 1000-fold different to the p-value derived from the logrank test (rs472660, rs4978394, rs3784780, rs6518956). For rs472660, the most significant model was recessive, and the heterozygote survival curves were close to one of the major homozygote curve, driving significance in this model in comparison to the minor homozygote. For rs4978394, the most significant logrank model

Figure 3-14 Kaplan Meier curves for rs472660 and rs3784780



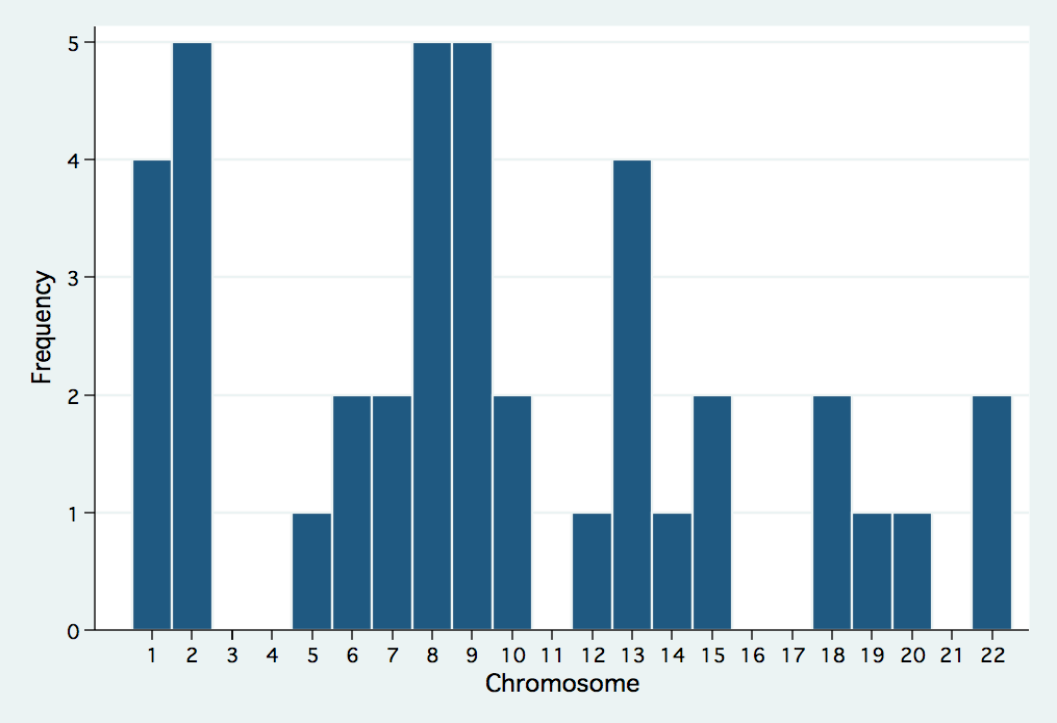
Kaplan Meier curve for rs472660 (left) illustrating how the BB and AB genotypes have very similar survival functions, while AA has a markedly worse survival, making the recessive model most significant in logrank testing (p=1.62e-09), but the trend across the three genotypes less so (p_{trend}=6.21e-08). For rs3784780 (right), the 'order' of the survival curves is not aligned with the coding for the trend test (AA=1, AB=2, BB=3), making the difference between the curves very significant (p=2.94e-06), but not the trend from AA to BB (p_{trend}=4.17e-01)

was genotypic, but the survival curves are suggestive of a more marked effect in the recessive model, and if the significance of the recessive model is compared to that of the logrank test, the discrepancy diminishes. For two SNPs, the most significant model was genotypic, and the survival curves were 'out of order' compared to the coding: for rs3784780, the order of decreasing survival was AB, BB, AA, while for rs6518956, this order was BB, AA, AB, rendering the trend much less strong than the observed difference between the survival functions (Table 3-7 and Figure 3-14).In the absence of bias, the distribution across the genome should be roughly even, with an expectation that the larger the chromosome, the more 'significant' SNPs should be located there. The

distribution of SNPs by chromosome, given in Figure 3-15, shows that the 40 SNPs taken forward were spread across the genome, although some chromosomes did not carry any: five of the included SNPs each are located on chromosome 2, 8, and 9, whereas there are none on chromosomes 3, 4, 11, 16, 17, and 21. The difference to the expected distribution based on chromosome size (3.5 SNPs on chromosome 1 to 0.7 SNPs on chromosome 22), however, is not significant (p=0.115).

Three SNPs were located in exons, 18 in introns, and 19 in intergenic regions. Two of the SNPs located in exons were non-synonymous base changes (rs10429965 and rs25689) while the third was synonymous base change (rs4776494). For the intergenic SNPs, the median distance to the nearest gene was 77kb, the nearest gene was 600bp, the furthest 1.3Mb.

Figure 3-15 Distribution of SNPs by chromosome

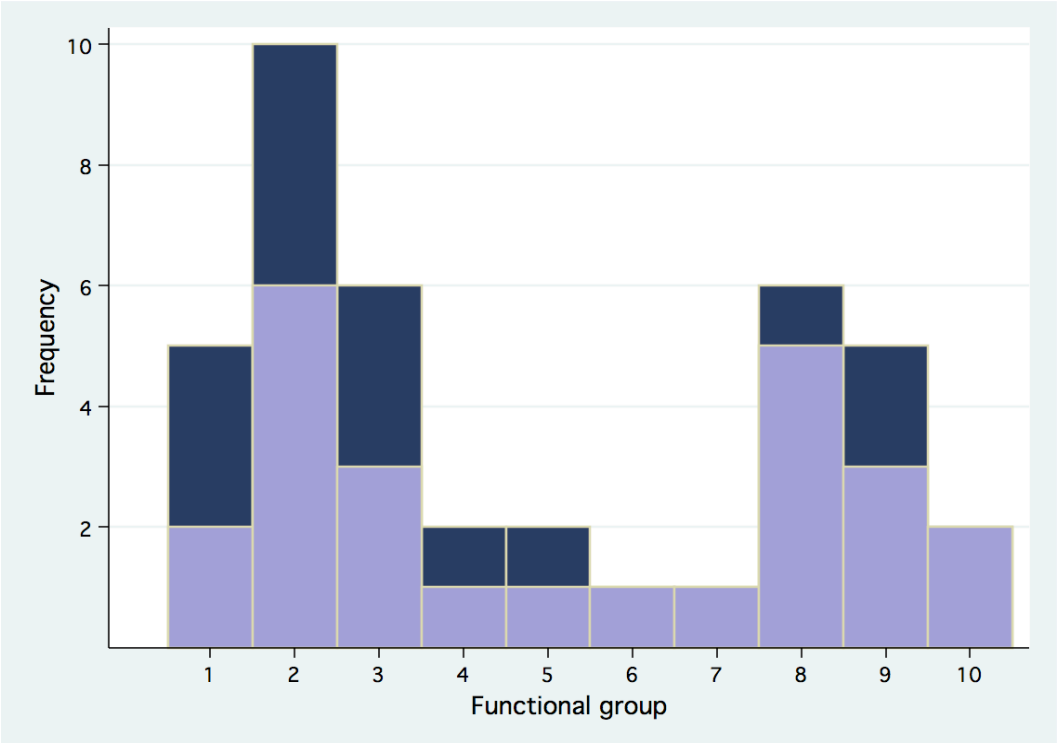


Histogram showing the frequency of significant association by individual chromosome. The distribution is not uniform, but not significantly different from the expected distribution if SNPs were purely picked at random.

There were a number of interesting candidate genes associated with the top 25 SNPs based on p-value alone, and, perhaps more expected as additional SNPs were chosen based on function, the list of 40 SNPs. The most prominent category was genes broadly involved in cell adhesion and motility (10 SNPs), followed by genes involved in cell signalling (6 SNPs) and transcriptional regulation (6 SNPs). These categories were also the most prominent when only the 25 SNPs chosen on statistical grounds were considered, with 6 SNPs in or near genes broadly involved in cell adhesion and motility, followed by transcriptional regulation (5 SNPs) and genes involved in cell signalling (4 SNPs). The cell

adhesion and motility group contains 2 SNPs which are associated with the same gene (rs1878632 and rs7600624, formin-like 2), while the transmembrane group contains at least one SNP in a gene encoding a transmembrane protein with unknown function, but thought to be tumour suppressor (rs1350308, *CSMD1*[297,298]). See Figure 3-16 for the distribution across putative gene function, indicating those that were chosen based on statistical significance and those chosen on biological rationale.

Figure 3-16 Distribution of functional categories across top 40 SNPs



1 - Cell cycling, 2 - Cell motility and adhesion, 3 - Cell signalling, 4 - DNA processing, 5 - Detoxification, 6 - Immune function, 7 - Redox system, 8 - Transcriptional regulator, 9 - Transmembrane protein, 10 – Unknown; lavender – SNPs chosen on statistical significance alone, navy – SNPs chosen on biological rationale.

3.7 Discussion

The data presented here show that using outcome in CRC as the complex trait in question, GWAS to determine novel markers of prognosis are feasible. Large patient series exist with good quality follow-up and sufficient germline DNA, high-density SNP typing platforms produce robust and reproducible genotypes, the statistical tools to analyse the data have been described, and the candidates selected for further validation largely fit with our current understanding of CRC biology. As was the case with GWAS successfully identifying CRC predisposition loci, SNPs described in phase 1 will require verification in further datasets, and two such series with 899 and 1338 patients have been identified: Scotland Phase 1, part of a GWAS for CRC risk[299], and PETACC 3, a clinical trial[142] and genotypes for selected SNPs are available from two population based cohorts.

The patients of the VICTOR trial had stage 2 and 3 CRC, perhaps the most interesting cohort to study for prognostic markers as the relapse rate in these stages is sufficiently high to warrant further intervention following surgery. In stage 1 CRC, the 5-year OS is in the region of 93%, making study difficult, and, unless the effect of a marker of poor prognosis carries a very large effect, intervention is unlikely to be acceptable to the patient. In stage 4 CRC, treatment is often driven by the general fitness of the patient, and prognostic markers are likely to less helpful in making treatment decisions.

The VICTOR trial design was that such that all patients had to have received standard surgery and (neo) adjuvant therapy as appropriate for the site and stage of disease. All patients receiving systemic therapy received 5-FU based chemotherapy, with only three patients also receiving a not otherwise specified novel agent. Subsequently, patients would be randomised to either rofecoxib for 2 or 5 years, or placebo for the same length of time. Due to the early closure of the trial following the worldwide withdrawal of rofecoxib due to excess cardiac events reported in the APPROVe study[300], the duration of intervention with rofecoxib was short (median 8.4 months, interquartile range 3.4-15.1 months). At present, the data regarding a possible protective effect of a COX-2 inhibitor are not known, but the short duration of therapy with rofecoxib makes it likely that this was not a confounder in this study, making the VICTOR cohort a homogeneous patient set with reliable ascertainment of disease status.

At the time of analysis, compared to reported rates, the event rate was lower than would be expected, 3-year DFS was 90.4% for stage 2, and DFS=82.8% for stage 3. The figures for stage 2 split into 93.0% for those who did not receive adjuvant therapy, and 89.2% for those who did. For stage 3, these figures are 69.6% and 83.2%, respectively. Reasons for this could be 1) the time from surgery to randomisation was substantial for some patients (median 7.7months, range 0.6 - 23.4months), although the majority of those delayed for more than 6months were patients with stage 3 disease who were randomised after the end of adjuvant chemotherapy; 2) sample collection was commenced 6 months after the trial closed, and while 66% of relapses were collected after the event, it is difficult to know how representative this set therefore is of the overall cohort as the final survival figures have not yet been published; 3) 11.8% of patients had not been followed for 3 years, and; and 4) the precise ascertainment of DFS is difficult due to differing approaches of treating physicians to the timing of cross-sectional imaging and biopsy to prove relapsed disease. However, the differences of outcome in stage 3 patients based on the adjuvant therapy is reassuring, as the size of the benefit associated with adjuvant 5-FU based chemotherapy is in keeping with published data[139]. The main implication for this study is that lower event rates decrease power, but as outlined in section 3.1, depending on allele frequency and risk conferred, power was sufficient to detect meaningful markers of prognosis.

In line with other GWAS and the technology reports on the performance of the Illumina SNP typing platform[270], genotyping was robust when applied to the samples of the VICTOR cohort. The vast majority of genotypes had good GenCall scores and, on a per SNP basis, good GenTrain scores. GenCall and GenTrain scores are measures validated internally by Illumina, and are not immediately amenable to cross checking, but visual inspection of cluster plots for individual SNPs confirmed the numerical assessment of genotyping, again with the overwhelming majority of randomly sampled plots showing good cluster separation (see Figure 3-3 for a representative plot). In addition, cluster plots for all SNPs with a call frequency of <98 .71% were inspected, and genotypes adjusted where necessary This cut off was set as the increase in call frequency following visual inspection had become negligible at this level. Lastly, internal (repeating the same analysis with the Illumina arrays) and external (generating genotypes by a different technology) reproducibility were high.

Based on the available survival data and genotypes, it was possible to select SNPs at significance levels seen in other GWAS, and a number of SNPs reached the proposed levels of genome-wide significance of p<1e-07. The strategy for the selection of SNPs to be taken into verification (Figure 3-10) was designed to avoid missing positive associations (type 2 error), therefore inclusion criteria were relatively lax: four models were analysed for their association with outcome, and the threshold for selection set at p<1e-04. To compensate for a possible over-inflation of associations (type 1 error), exclusion criteria were much more stringent: the significance level for exclusion for deviation from HWE, as a measure of potential confounding by hidden genetic substructure, was set at p=0.01, much higher than the level of genome-wide significance assumed for outcome. Only cases were allowed to deviate from HWE at lower p-values as this could be representative of the underlying, true association, as the HWE assumption may not hold if samples are chosen based on relapse[301].

Further, SNPs with low minor homozygote frequencies were excluded, as 'chance' variation in small groups appeared to have a disproportionately large effect on the resulting effect size and p-value in the logrank model (Figure 3-9), although this did not affect λ to an extent that would have suggested a hidden population substructure. The net effect of the strategy employed here was to consider only SNPs with a MAF of >0.12, effectively excluding SNPs for which this study would not have had sufficient power. It did, however, exclude some SNPs that showed highly significant p-values, and in the case of the most significant SNP included for verification (rs472660), in a region spanning 40Kb, two SNPs have much lower p-values (rs585071 and rs678040, in both cases p=4.35E-14) driven by minor homozygote counts of 0 in the control group and 5 in the relapsed group. A further SNP (rs667660, p=1.59e-08) had counts of 1 and 5,

respectively. All three would have been included if the MAF cut-off had been set to 0.05 (Figure 3-12).

One concern was that the strategy of excluding low minor homozygote groups based on the control group might have biased against SNPs strongly associated with relapse. The level used is equivalent to 17-18 in the minor homozygote group for the overall cohort, and at that level, only four SNPs with a minor homozygote count of 9 in the control group would have been included in addition to the SNPs eligible for verification as described above.

The predominant model associated with a significant effect was the recessive model; mostly because of the effect of small numbers on the logrank test for equality of the survivor function. This is known[302], and due to the difficulty in estimating the asymptotic test value (one that would be derived if an infinite number of subjects were available) by only a small number of observations. This problem will complicate any GWAS for prognosis when compared to those GWAS for risk analysing any model other than the genotypic one. In the case of relatively small numbers of events, as was the case in the VICTOR cohort at the time of SNP selection, tests of proportion (χ^2 or Fisher's exact test) give very similar estimates of significance to a time-to-event test such as the logrank test[303], but avoid the overestimation of χ^2 values at the extremes, and a trend test such as Cox regression of all genotypes grouped separately also appear more robust in this respect and should be the analysis of choice if time-to-event data is to be included.

The stipulation that SNPs had to be in HWE could have biased against SNPs that carry prognostic information, but are also strongly associated with CRC risk, thus distorting the HWE in the VICTOR cohort. This is unlikely to have been a factor, as most of the patients in the VICTOR cohort, on the grounds of being a population based cohort, are almost certain to have had sporadic CRC. Further, as the data in Chapter 6 demonstrates, the SNPs with the strongest link to CRC risk were in HWE and did not show an association with prognosis. It is therefore unlikely that there are other SNPs that are associated with CRC sufficiently strongly to have a distorted HWE.

Fifteen SNPs of the 103 SNPs making the significance cut-off were selected based on biological rationale, introducing an element of hypothesis driven research. It was hoped that by choosing SNPs that had a higher biological rationale, they might be more likely to replicate. Given the paucity of validated hypothesis driven markers of prognosis, discussed in section 0, only SNPs meeting the initial p-value cut-off were considered. These 15 SNPs represent a compromise between the need to validate more markers than meet the most stringent significance thresholds to minimise the type 2 error and the cost associated

with validation of many hundreds of markers, both financially and in terms of DNA usage in patient series with good quality outcome data but limited amounts of DNA.

Even without the 15 SNPs chosen in this fashion, the putative functional distribution of the nearest gene to the SNPs chosen based on p-value suggests that several SNPs in this group also present a plausible biological rationale: rs4649314 lies 50Kb downstream of mixed lineage kinase 4 (*MLK4*), a mitogen-activated protein kinase kinase kinase (MAP3K) on chromosome 1, and known activator of the JUN N-terminal kinase (*JNK*) and p38 mitogen activated protein kinase (*MAPK*) pathways[304]. The interplay of these pathways is involved in a plethora of cellular homeostatic functions (e.g. proliferation, apoptosis, migration[305]), and alterations in the MAPK pathway have been implicated in colorectal carcinogenesis[306], although there is no evidence that *MLK4* specifically is a factor in CRC. rs7556894 is located 15Kb downstream from Actin-related protein 2 (*ARP2*), a member of the ARP2/3 complex involved in branching actin filaments and essential in cell shape and motility[307], and actin filaments interact with LIM kinase during the metastatic process[290]. rs5997921, chosen on biological rationale, lies intronic in LIM2 kinase and together, these two SNPs could implicate the metastatic pathway in the determination of prognosis. rs6972789 lies 350Kb downstream of stromal antigen 3-like 4, which appears down-regulated in CRC. It is conceivable that rs6972789 could lie in a long range enhancer element influencing the expression of stromal antigen3-like 4 as at least 50% of such elements are more than 250Kb from 'their' gene[308].

One functional category not represented in the 40 SNPs was pharmacogenetic markers. This was to be expected as the analyses performed did not take into account treatment (given to just under two thirds of patients in the VICTOR trial), and therefore predictive markers were unlikely to be identified. Looking at SNPs in genes (and their neighbouring regions) involved in the metabolism of 5-FU did not reveal any associations with Bonferroni corrected $p<0.05$ (data not presented in this thesis). For a full functional annotation of the top 25 SNPs chosen on p-value alone see Table 10-2, even if at this stage, they are merely interesting candidates, without a proven association with outcome.

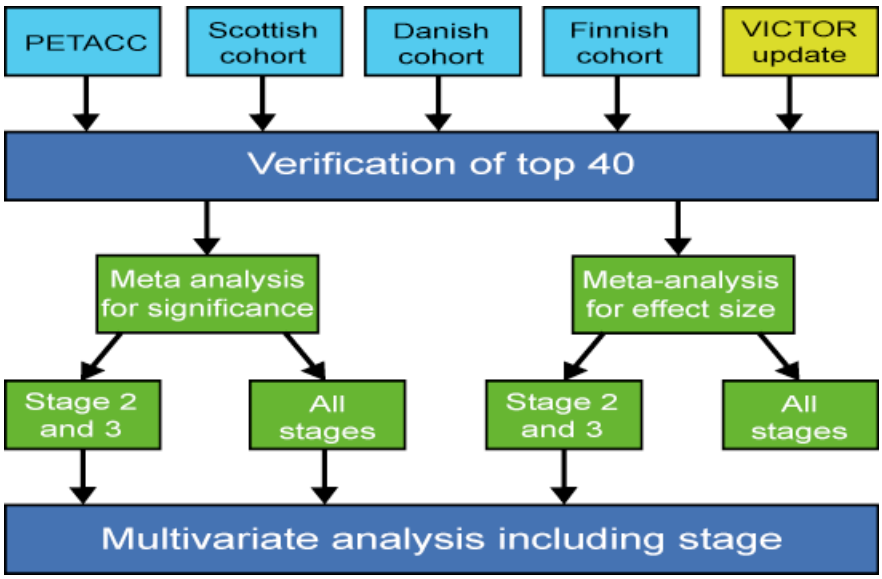
The data presented in this chapter support the view that GWAS for outcome in CRC are feasible, that the data generated is robust at the genotyping level and plausible in the survival analysis. Despite the drawbacks discussed, the top 40 SNPs look interesting and their attempted verification is presented in Chapter 4.

Chapter 4

GWAS for prognostic markers: Verification phase

Replication of the findings in chapter 3 was attempted in two cohorts (PETACC and Scotland Phase 1), with genotypes available for selected SNPs in two further cohorts. All genotypes were analysed for the model for which the SNP in question had been most significant in Phase 1. Unlike phase 1 described in chapter 3, in this chapter, all p-values are trend as they pertain to the HR and CI to be meta-analysed. In the case of allelic OR, the p-value is derived using Fisher's exact test. Genotypes are represented according to the Illumina convention.

Figure 4-1 Flowchart of verification analysis



The verification was attempted in two phases: for significance using the raw ES and CI, then to derive an estimate of effect size, using adjusted data for the VICTOR cohort (see section 4.3).

The main aim of the verification phase is to derive an estimate of statistical significance (section 4.2) , although an estimate of the size of the effect can also be given (section 4-3), and the flowchart of the verification analysis is shown in Figure 4-1.

4.1 Outcome in individual series

In addition to the PETACC cohort identified at the outset of the discovery phase, a further cohort with genotype data for all but one SNP was identified (Scotland phase 1). Genotypes for a small proportion of SNPs were also available from a Danish cohort (10 SNPs) and a Finnish cohort (1 SNP).

Table 4-1 Summary of cohorts

Cohort	Relapse	Non-relapse	Total	Number of SNPs
VICTOR	184	763	947	40
Scotland	289	610	899	39
PETACC	386	865	1253	33
Denmark	207	392	599	10
Finland	322	638	960	1
Total	1388	3268	4656	

Relapsed and non-relapsed patients for each cohort used in the verification phase, and number of SNPs typed in each cohort.

4.1.1 PETACC

Collaborators at the EORTC made available genotyping data for 33 of the 40 selected SNPs in 1338 patients with stage 2 and 3. Survival data were progression-free survival, with mature 4-year follow-up and 386 relapsed patients. Genotyping was performed in the laboratory of Dr Diether Lambrechts at the University in Leuven, Belgium; the survival analysis was performed by Dr Mauro Delorenzi at the Swiss Institute of Bioinformatics, Lausanne.

Genotypes were generated using MassARRAY® iPLEX Gold single-base extension technology (Sequenom, CA) according to the manufacturer's protocol. Of the 40 SNPs put forward from the discovery phase, seven SNPs did not perform sufficiently well during the initial primer design and were not included in the genotyping (rs10842099, rs1350308, rs1526884, rs4776494, rs5997921, rs672757, rs7912136). The SNPs were genotyped in two batches, of 18 and 17 SNPs; there was insufficient DNA to type the remaining SNPs with redesigned primers.

Samples with >5 genotypes missing were deemed to have failed. Genotyping did not performed as well as the Illumina arrays, with 85 samples excluded from the first run, and 44 from the second, leaving 1253 and 1294 samples, respectively. After exclusion of the failed samples, the median call frequency was 98.7% (range 90.1%-100.0%), showing the genotyping of the included samples was robust. Two SNPs from the first run were

repeated due to a low call frequency in one case (rs3752261, 80.2%) and poor clustering in the other (rs1924597). Both these SNPs performed well in the second run. After exclusion of the globally failed samples, the call rates between the two runs did not differ significantly (p=0.478).

Table 4-2 Genotypes and hazard ratios for 33 SNPs in PETACC 3

	Case			Control					
SNP	AA	AB	BB	AA	AB	BB	Model	ES (95% CI)	p-value
rs472660	8	187	631	9	72	279	recessive	2.17 (1.12-4.21)	2.15e-02
rs7556894	51	352	460	33	148	205	recessive	1.41 (0.98-2.01)	6.05e-02
rs2589183	39	300	525	22	131	233	allelic	1.05 (0.85-1.28)	6.77e-01
rs764372	17	220	628	10	99	277	recessive	1.28 (0.69-2.41)	4.34e-01
rs6972789	258	410	169	108	191	77	recessive	0.99 (0.77-1.27)	9.17e-01
rs7007146	213	387	164	91	168	65	recessive	0.91 (0.69-1.20)	5.03e-01
rs745888	71	332	451	31	146	205	recessive	0.95 (0.66-1.37)	7.77e-01
rs4978394	583	162	19	246	76	11	genotypic	1.10 (0.89-1.35)	3.67e-01
rs3784780	25	227	550	11	103	250	genotypic	1.01 (0.83-1.22)	9.53e-01
rs10510044	727	78	2	327	25	1	recessive	0.90 (0.13-6.38)	9.13e-01
rs712082	28	255	582	14	105	267	recessive	1.09 (0.64-1.85)	7.58e-01
rs11788150	121	376	338	53	168	151	recessive	0.95 (0.71-1.27)	7.34e-01
rs4715476	563	238	26	244	113	10	genotypic	1.02 (0.85-1.23)	8.28e-01
rs4649314	93	384	342	56	173	141	recessive	1.31 (0.98-1.74)	6.40e-02
rs9533457	390	369	74	181	160	31	recessive	0.96 (0.66-1.39)	8.26e-01
rs1571583	56	277	509	25	127	220	recessive	1.01 (0.67-1.51)	9.80e-01
rs567564	72	382	370	35	148	185	recessive	1.05 (0.74-1.49)	7.86e-01
rs7866165	85	355	420	44	145	196	recessive	1.11 (0.81-1.51)	5.33e-01
rs437171	45	258	520	21	98	249	recessive	1.05 (0.68-1.63)	8.31e-01
rs3752261	591	247	25	257	114	12	recessive	1.05 (0.59-1.86)	8.73e-01
rs1924597	634	188	11	297	71	4	allelic	0.82 (0.63-1.08)	1.75e-01
rs1822917	528	281	29	228	129	15	recessive	1.25 (0.74-2.09)	4.04e-01
rs9514816	19	175	628	6	75	276	allelic	0.93 (0.71-1.22)	6.38e-01
rs6518956	280	412	138	128	183	58	genotypic	0.96 (0.83-1.11)	5.92e-01
rs1878632	262	423	149	110	177	85	recessive	1.26 (0.99-1.61)	6.10e-02
rs7600624	609	237	20	264	105	16	allelic	1.14 (0.91-1.42)	2.68e-01
rs9315425	16	181	639	7	79	286	recessive	0.99 (0.47-2.10)	9.87e-01
rs10429965	627	193	14	278	85	6	recessive	0.97 (0.43-2.17)	9.35e-01
rs1824476	82	362	391	39	163	168	allelic	1.05 (0.87-1.26)	6.02e-01
rs1438620	288	371	160	117	163	84	dominant	1.14 (0.92-1.43)	2.32e-01
rs2514841	180	369	222	82	158	102	recessive	1.05 (0.82-1.34)	7.09e-01
rs1034116	11	205	650	7	92	286	allelic	1.06 (0.83-1.36)	6.56e-01
rs25689	37	298	530	25	121	239	allelic	1.04 (0.85-1.28)	7.13e-01

The genotypes for cases and control in PETACC 3. The effect size (ES) for the allelic model is OR, when the p-value is derived using Fisher's exact test, for all other models, the ES is HR, and the p-value is the p_{trend} from Cox regression. The genetic model tested is the same as in the discovery phase in VICTOR. SNPs with an ES in the same direction as the VICTOR cohort are indicated in light grey, those not included in further analysis are indicated in dark grey.

Three SNPs had a significantly different genotype distribution at the 0.01 level compared to the VICTOR cohort: rs2514841 (p=9.00e-03), rs7866165 (p=4.74e-07), and

rs10510044 (p=5.02E-100). The data from PETACC for the latter SNP was excluded from further analysis, as the analysis of the raw genotyping data (spectral frequencies of the Sequenom) showed this to be least reliable of the SNPs typed. In addition, two SNPs were not in HWE: rs1438620 (p=0.006) and rs437171 (p=0.004), these were also excluded, leaving genotypes for 30 SNPs for meta-analysis. For 45 samples, the survival analysis could not be reliably performed, and genotypes for these were excluded from logrank testing and Cox regression. The final distribution of the genotypes for all 33 SNPs is given in Table 4-2, and excluded SNPs are indicated.

Only one SNP had p<0.05 in the PETACC cohort (rs472660), but 14 SNPs had an ES of greater than 1.05 (or less than 0.95) going in the same direction as that found in the VICTOR cohort (rs472660, rs7556894, rs764372, rs4978394, rs712082, rs4649314, rs567564, rs7866165, rs3752261, rs1924597, rs1822917, rs1878632, rs7600624, rs1034116).

4.1.2 Scottish cohort

Collaborators in Edinburgh made available clinical and genotyping data for 899 patients of all stages, this cohort is a subset of the 1,012 patients described under Scotland Phase 1 in section 2.1.3.3, part of a GWAS for CRC risk[299]. Survival data were CRC specific survival. After a median follow-up of 69.3 months there were 289 CRC specific deaths. There were 561 patients with stage 2 and 3 CRC, and 130 CRC specific deaths.

Genotypes were available for 39 SNPs, and all SNPs were typed on Illumina arrays according to the manufacturer's protocol; rs1526884 failed genotyping. For the remaining SNPs, genotyping performed well, with all but four SNPs having been typed successfully for all patients. For these four SNPs, the call frequency was >99.1% with one or two genotypes missing for three SNPs, and eight for one SNP (rs672757). One SNP had a significantly different genotype distribution between the VICTOR cohort and Scotland Phase 1 (rs1571583, p=0.001), although the allele distribution was not significantly different once adjusted for multiple comparisons (p_{all}=0.016). The genotype distribution for the other SNPs was not significantly different at the 0.01 level. All SNPs in Scotland Phase 1 were in HWE. The distribution of the genotypes is given in Table 4-3.

Only one of the 39 SNPs had p≤0.05 in the Scottish cohort (rs1824476), and 15 SNPs had an ES of greater than 1.05 (or less than 0.95) going in the same direction as that found in the VICTOR cohort (rs764372, rs7007146, rs712082, rs11788150, rs4649314, rs1571583, rs567564, rs437171, rs1924597, rs1878632, rs7600624, rs5997921, rs1824476, rs1438620, rs2514841).

For a comparison with the PETACC data, no SNP in the Scottish cohort reached p<0.05 with an ES going in the same direction in both cohorts. There were seven SNPs with an

ES of greater than 1.05 (or less than 0.95) going in the same direction as that found in PETACC (rs764372, rs712082, rs4649314, rs567564, rs1924597, rs1878632, rs7600624).

Table 4-3 Genotypes and hazard ratios for 39 SNPs in Scotland Phase 1

	Case			Control					
SNP	AA	AB	BB	AA	AB	BB	Model	ES (95% CI)	p-value
rs472660	16	130	464	6	57	226	recessive	0.80 (0.36-1.80)	5.95e-01
rs7556894	44	225	341	19	110	160	recessive	0.90 (0.56-1.43)	6.54e-01
rs2589183	33	192	385	15	91	183	allelic	0.99 (0.77-1.26)	9.51e-01
rs764372	16	140	454	9	69	211	recessive	1.19 (0.61-2.31)	6.08e-01
rs672757	121	291	192	51	141	95	allelic	0.93 (0.76-1.14)	5.06e-01
rs7912136	308	245	57	141	124	24	recessive	0.89 (0.59-1.36)	5.95e-01
rs6972789	181	292	137	86	142	61	recessive	0.95 (0.72-1.26)	7.15e-01
rs7007146	170	306	134	64	155	70	recessive	1.06 (0.81-1.39)	6.56e-01
rs745888	50	221	339	19	122	148	recessive	0.90 (0.56-1.43)	6.41e-01
rs4978394	453	142	15	208	76	5	genotypic	1.04 (0.82-1.30)	7.61e-01
rs3784780	14	176	420	8	82	199	genotypic	1.01 (0.81-1.26)	9.20e-01
rs10510044	312	247	51	141	128	20	recessive	0.81 (0.51-1.27)	3.50e-01
rs712082	25	177	408	16	76	197	recessive	1.22 (0.73-2.01)	4.48e-01
rs11788150	76	275	259	40	136	112	recessive	1.11 (0.80-1.56)	5.24e-01
rs10842099	339	225	46	154	119	16	allelic	1.01 (0.80-1.26)	9.54e-01
rs4715476	383	200	26	182	96	10	genotypic	0.99 (0.81-1.22)	9.59e-01
rs4649314	77	290	243	51	135	102	recessive	1.33 (0.98-1.80)	6.28e-02
rs1350308	102	286	222	47	143	99	allelic	1.04 (0.85-1.27)	7.58e-01
rs9533457	326	239	45	155	120	14	recessive	0.69 (0.40-1.18)	1.80e-01
rs1571583	34	185	391	18	90	181	recessive	1.11 (0.69-1.78)	6.76e-01
rs567564	52	255	303	27	102	160	recessive	1.11 (0.74-1.64)	6.18e-01
rs7866165	47	240	323	23	115	151	recessive	1.02 (0.67-1.56)	9.31e-01
rs437171	18	190	402	10	76	203	recessive	1.20 (0.64-2.26)	5.64e-01
rs3752261	414	169	27	181	100	8	recessive	0.68 (0.34-1.37)	2.82e-01
rs1924597	440	158	12	219	62	8	allelic	0.89 (0.67-1.18)	4.73e-01
rs1822917	374	207	29	166	109	14	recessive	1.01 (0.59-1.73)	9.66e-01
rs9514816	9	151	450	6	59	224	allelic	0.87 (0.65-1.17)	3.74e-01
rs6518956	218	298	94	107	134	48	genotypic	0.97 (0.82-1.14)	6.94e-01
rs1878632	214	296	100	96	140	53	recessive	1.11 (0.83-1.50)	4.79e-01
rs7600624	442	158	10	200	84	5	allelic	1.14 (0.87-1.49)	3.60e-01
rs9315425	13	155	442	3	74	212	recessive	0.52 (0.17-1.62)	2.58e-01
rs10429965	468	133	9	216	69	4	recessive	0.84 (0.31-2.25)	7.29e-01
rs5997921	13	158	439	7	51	231	recessive	1.07 (0.50-2.26)	8.64e-01
rs1824476	60	246	304	19	109	161	allelic	0.80 (0.64-1.00)	5.00e-02
rs4776494	441	153	16	206	79	4	genotypic	0.99 (0.79-1.24)	9.15e-01
rs1438620	208	297	105	89	143	57	dominant	1.16 (0.90-1.49)	2.48e-01
rs2514841	125	293	192	51	154	84	recessive	0.85 (0.63-1.15)	2.87e-01
rs1034116	8	150	452	3	68	218	allelic	0.93 (0.69-1.25)	6.57e-01
rs25689	20	231	359	14	83	192	allelic	0.83 (0.65-1.07)	1.56e-01

The genotypes for cases and control in Scotland Phase 1. The effect size (ES) for the allelic model is OR, when the p-value is derived using Fisher's exact test, for all other models, the ES is HR, and the p-value is the p_{trend} from Cox regression. The genetic model tested is the same as in the discovery phase in VICTOR. SNPs with an ES in the same direction as the VICTOR cohort are indicated in light grey.

4.1.3 VICTOR update

Following the discovery phase, genotypes for a further 16 non-relapsed patients became available, as well as updated survival data. Follow-up lengthened by 261 days at the new

Table 4-4 Genotypes and hazard ratios for 40 SNPs in VICTOR

	Case			Control					
SNP	AA	AB	BB	AA	AB	BB	Model	ES (95% CI)	p-value
rs472660	12	43	129	10	169	584	recessive	3.95 (2.20-7.09)	4.43e-06
rs7556894	26	63	95	38	289	435	recessive	3.03 (2.00-4.59)	1.83e-07
rs2589183	4	36	144	44	248	470	allelic	0.48 (0.34-0.67)	8.70e-06
rs764372	11	50	123	11	195	557	recessive	3.79 (2.06-6.99)	1.91e-05
rs672757	45	97	42	105	379	279	allelic	1.64 (1.31-2.07)	2.63e-05
rs7912136	93	59	32	382	323	57	recessive	2.36 (1.61-3.45)	1.07e-05
rs6972789	42	73	69	216	396	151	recessive	2.23 (1.65-3.01)	1.46e-07
rs7007146	39	80	65	213	385	165	recessive	1.85 (1.37-2.51)	6.75e-05
rs745888	25	71	88	44	316	403	recessive	2.35 (1.54-3.58)	7.44e-05
rs4978394	118	54	12	568	182	13	genotypic	1.58 (1.24-2.03)	2.66e-04
rs3784780	11	36	137	14	213	535	genotypic	0.99 (0.75-1.32)	9.51e-01
rs10510044	76	75	33	401	300	62	recessive	2.26 (1.55-3.30)	2.18e-05
rs712082	20	45	118	25	227	509	recessive	2.87 (1.80-4.57)	8.77e-06
rs11788150	39	72	73	75	327	360	recessive	2.26 (1.59-3.23)	6.44e-06
rs10842099	119	58	7	372	317	74	allelic	0.56 (0.42-0.73)	2.05e-05
rs4715476	102	69	13	521	223	18	genotypic	1.67 (1.31-2.13)	3.11e-05
rs1526884	121	49	13	381	311	71	dominant	0.53 (0.39-0.73)	6.05e-05
rs4649314	37	72	73	80	367	314	recessive	2.05 (1.43-2.95)	9.51e-05
rs1350308	50	86	48	119	344	299	allelic	1.65 (1.32-2.08)	1.96e-05
rs9533457	88	70	26	413	307	43	recessive	2.24 (1.48-3.40)	1.38e-04
rs1571583	21	72	91	27	294	442	recessive	2.71 (1.72-4.26)	1.79e-05
rs567564	34	75	75	59	324	379	recessive	2.43 (1.67-3.53)	2.99e-06
rs7866165	19	63	102	29	290	444	recessive	2.68 (1.67-4.31)	4.91e-05
rs437171	16	58	110	18	234	511	recessive	2.99 (1.79-5.00)	2.85e-05
rs3752261	109	63	12	516	231	16	recessive	2.90 (1.61-5.21)	3.66e-04
rs1924597	165	16	3	550	196	17	allelic	0.36 (0.23-0.56)	8.88e-07
rs1822917	109	52	22	437	291	35	recessive	2.39 (1.53-3.74)	1.27e-04
rs9514816	7	59	118	9	153	601	allelic	1.96 (1.45-2.65)	2.73e-05
rs6518956	53	114	17	279	341	141	genotypic	0.98 (0.80-1.20)	8.34e-01
rs1878632	61	73	50	263	375	125	recessive	1.84 (1.33-2.55)	2.41e-04
rs7600624	119	54	11	576	174	13	allelic	1.73 (1.29-2.31)	3.82e-04
rs9315425	10	39	135	13	199	551	recessive	2.95 (1.56-5.59)	8.84e-04
rs10429965	140	33	11	592	163	8	recessive	4.59 (2.49-8.46)	1.06e-06
rs5997921	9	42	133	12	163	587	recessive	2.89 (1.48-5.65)	1.93e-03
rs1824476	9	62	113	67	331	365	allelic	0.63 (0.48-0.83)	8.35e-04
rs4776494	119	57	8	562	188	12	genotypic	1.55 (1.20-2.02)	8.93e-04
rs1438620	36	99	49	256	353	153	dominant	2.00 (1.39-2.88)	2.00e-04
rs2514841	20	110	53	166	396	201	recessive	0.46 (0.29-0.74)	1.14e-03
rs1034116	9	58	117	13	162	587	allelic	1.85 (1.38-2.48)	7.46e-05
rs25689	16	75	93	25	269	469	allelic	1.55 (1.20-2.00)	1.05e-03

The genotypes for cases and controls in VICTOR after the addition of a further 16 samples. The effect size (ES) is the OR for the allelic model, when the p-value is derived using Fisher's exact test, for all other models, the ES is HR, and the p-value is the p_{trend} from Cox regression.

cut-off date of 29 April 2009. The median follow-up was 51.2 months, SD ± 15.1 months for all patients, and 54.6 months ± 11.3 months for non-relapsed patients.

The 16 additional samples performed well on genotyping, with a median call rate of 99.84% (range 99.54%-99.95%). After the addition of the16 samples, all 40 SNPs remained in HWE. In most cases, the additional samples decreased statistical significance, although for ten SNPs, the association with outcome appeared more statistically significant after the update (rs7556894, rs6972789, rs712082, rs11788150, rs4715476, rs1571583, rs567564, rs437171, rs1924597, rs10429965). The resulting genotype distributions and HR are given in Table 4-4.

4.1.4 Genotypes for 10 SNPs in the Danish cohort

Genotypes were available for 10 SNPs (rs1034116, rs1350308, rs1571583, rs1822917, rs1924597, rs2589183, rs3784780, rs6518956, rs7556894, rs7600624) in a Danish cohort of 599 patients with all stages of CRC, of whom 401 had stage 2 and 3 CRC. The SNPs were typed using KASPar chemistry; the primers are given in Table 10-8, and the conditions used in Table 10-9, both in Appendix C. Except for one SNP (rs1571583), with a call frequency of 75.6%, genotyping performed well with a median call frequency of 98.0% (range 95.8%-98.3%). As the low call rate for rs1571583 was driven by the poor performance of two 96-well plates, these plates were excluded, after which the call frequency was 95.8%.

Table 4-5 Genotypes and hazard ratios for 10 SNPs in Danish cohort

	Case			Control					
SNP	AA	AB	BB	AA	AB	BB	Model	ES (95% CI)	p-value
rs7556894	14	72	114	23	151	215	recessive	1.27 (0.74-2.20)	1.50E-01
rs2589183	8	60	133	11	113	263	allelic	0.91 (0.66-1.24)	1.85E-01
rs3784780	10	56	132	5	111	267	genotypic	1.16 (0.90-1.51)	2.56E-01
rs1350308	33	102	65	55	191	141	allelic	0.88 (0.69-1.12)	3.02E-01
rs1571583	10	65	75	16	107	180	recessive	1.31 (0.69-2.49)	3.82E-01
rs1924597	150	47	5	295	88	4	allelic	0.86 (0.61-1.23)	4.09E-01
rs1822917	128	56	14	217	137	23	recessive	1.12 (0.65-1.93)	4.11E-01
rs6518956	61	106	26	96	224	61	genotypic	0.85 (0.68-1.06)	5.35E-01
rs7600624	138	57	3	269	108	9	allelic	1.03 (0.74-1.43)	6.88E-01
rs1034116	5	52	144	10	78	300	allelic	0.79 (0.56-1.12)	8.56E-01

Genotypes and HR in the Danish cohort. The effect size (ES) for the allelic model is OR, when the p-value is from Fisher's exact test, for all other models, the ES is HR, and the p-value is the p_{trend} from Cox regression. The genetic model tested is the same as in the discovery phase in VICTOR. SNPs with an ES in the same direction as the VICTOR cohort are indicated in light grey. rs6518956 was excluded from further analysis as it was not in HWE, and significantly differently distributed to the VICTOR cohort (shaded dark grey).

All SNPs were in HWE at the 0.05 level except for one SNP (rs6518956, p_{HWE}=3.28e-04). This SNP also had a significantly different genotype distribution compared with the VICTOR cohort and was therefore excluded, leaving nine SNPs for further analysis

(Table 4-4). Only one SNP reached statistical significance at the 0.05 level (rs3784780), and two (rs7556894, rs1571583) had an ES of greater than 1.05 (or less than 0.95) going in the same direction as that found in the VICTOR cohort. For rs7556894 this was also the case for the comparison with the PETACC cohort.

4.1.5 rs4649314 in the Finnish cohort

Genotypes for rs4649314 were available in a further verification cohort comprising 962 Finnish patients of all stages, a subset of the cohort described in section 2.1.4. Overall survival was available for 959 patients with a median follow-up of 39.1 months; there were 638 deaths. There were 655 patients with stage 2 and 3 CRC. The genotypes were generated using KASPar allele specific PCR and the primers are given in Table 10-8, and the conditions used in Table 10-9, both in Appendix C. Genotyping performed well with a call rate of 97.1%, and the SNP was in HWE in this cohort. The HR was of a slightly smaller magnitude than in the other three cohorts where it had been typed, but consistent with an effect in the same direction (HR=1.08, 95% CI 0.89-1.31, p=4.59e-01).

4.2 Meta-analysis for significance

Meta-analysis was performed to determine the overall significance level for every SNP for which there was data from two or more cohorts. rs1526884 was excluded as this had only been typed successfully in the VICTOR cohort; data for SNPs that deviated from HWE (rs1438620 and rs437171 in PETACC; rs6518956 in the Danish cohort) or had a significantly different genotype distribution from VICTOR (rs10510044 in PETACC) were also not included. The meta-analyses were for the same genetic model as initially detected in the VICTOR cohort, using a fixed effects model. All effect sizes were based on the latest available survival data and are given in Table 4-2, Table 4-3, Table 4-4, Table 4-4, and section 4.1.5.

4.2.1 Meta-analysis for all stages

This analysis included all patients for whom genotypes were available, i.e. stage 1 to 4 patients from the Scottish, Danish, and Finnish cohorts. Overall, there were 10 SNPs typed in four cohorts, 24 SNPs in three cohorts, and 5 SNPs in two cohorts.

The most significant SNP was rs472660, intronic to *CYP3A43* protein on chromosome 7, with p=4.01e-05, but this did not reach genome-wide significance (p<1e-07). This SNP had data available from three cohorts; with PETACC showing a similar ES to VICTOR (HR=2.17, 95% CI 1.12-4.21, p=2.15e-02), but not the Scottish cohort (HR=0.80, 95% CI 0.36-1.80, p=5.95e-01). The summary ES for each SNP is given in Table 4-6.

Table 4-6 Summary ES and 95%CI for meta-analysis

SNP	Cohorts	Model	p-value	ES (95% CI)	I^2	p_Q
rs472660	3	recessive	4.01e-05	2.24 (1.53-3.30)	79.5	7.55e-03
rs7556894	4	recessive	7.38e-05	1.54 (1.25-1.92)	81.2	1.15e-03
rs437171	2	recessive	3.05e-04	2.08 (1.40-3.10)	79.2	2.83e-02
rs4649314	4	recessive	3.21e-04	1.27 (1.12-1.45)	69.2	2.10e-02
rs1878632	3	recessive	4.97e-04	1.34 (1.13-1.57)	62.7	6.84e-02
rs712082	3	recessive	7.70e-04	1.64 (1.23-2.19)	78.5	9.50e-03
rs764372	3	recessive	9.84e-04	1.85 (1.28-2.66)	75.9	1.58e-02
rs10429965	3	recessive	1.00e-03	2.08 (1.35-3.22)	84.8	1.41e-03
rs1924597	4	allelic	1.27e-03	0.77 (0.65-0.90)	76.1	5.70e-03
rs567564	3	recessive	1.81e-03	1.41 (1.14-1.74)	83.9	2.02e-03
rs1438620	2	dominant	2.26e-03	1.38 (1.12-1.69)	82.9	1.57e-02
rs7600624	4	allelic	3.65e-03	1.22 (1.07-1.40)	57.2	7.14e-02
rs7912136	2	recessive	3.83e-03	1.52 (1.14-2.01)	91.2	7.73e-04
rs1822917	4	recessive	5.27e-03	1.43 (1.11-1.85)	61.4	5.12e-02
rs1571583	4	recessive	5.67e-03	1.40 (1.10-1.77)	74.1	8.91e-03
rs10510044	2	recessive	7.63e-03	1.48 (1.11-1.98)	91.5	5.99e-04
rs4978394	3	genotypic	7.82e-03	1.19 (1.05-1.36)	71.9	2.84e-02
rs10842099	2	allelic	1.11e-02	0.80 (0.67-0.95)	90.5	1.14e-03
rs6972789	3	recessive	1.17e-02	1.23 (1.05-1.44)	90.7	2.22e-05
rs11788150	3	recessive	1.21e-02	1.27 (1.05-1.53)	86.3	6.85e-04
rs5997921	2	recessive	1.53e-02	1.86 (1.13-3.06)	73.4	5.24e-02
rs7866165	3	recessive	1.67e-02	1.31 (1.05-1.64)	82.1	3.79e-03
rs1824476	3	allelic	2.06e-02	0.86 (0.76-0.98)	79.6	7.38e-03
rs672757	2	allelic	2.31e-02	1.19 (1.02-1.39)	92.6	2.45e-04
rs4715476	3	genotypic	2.96e-02	1.14 (1.01-1.29)	84.2	1.78e-03
rs1350308	3	allelic	3.59e-02	1.15 (1.01-1.31)	87.0	4.69e-04
rs4776494	2	genotypic	3.60e-02	1.20 (1.01-1.42)	85.0	9.89e-03
rs7007146	3	recessive	4.57e-02	1.18 (1.00-1.39)	84.0	1.90e-03
rs9315425	3	recessive	6.12e-02	1.53 (0.98-2.39)	77.4	1.21e-02
rs745888	3	recessive	7.00e-02	1.25 (0.98-1.58)	84.3	1.74e-03
rs3752261	3	recessive	8.69e-02	1.36 (0.96-1.94)	81.7	4.19e-03
rs2589183	4	allelic	9.67e-02	0.90 (0.79-1.02)	81.2	1.14e-03
rs2514841	3	recessive	1.10e-01	0.87 (0.72-1.03)	78.5	9.55e-03
rs9514816	3	allelic	1.16e-01	1.14 (0.97-1.35)	88.8	1.37e-04
rs9533457	3	recessive	1.36e-01	1.20 (0.94-1.54)	85.9	8.42e-04
rs1034116	4	allelic	1.40e-01	1.12 (0.96-1.29)	82.0	8.18e-04
rs25689	3	allelic	2.22e-01	1.09 (0.95-1.24)	83.4	2.45e-03
rs6518956	4	genotypic	2.30e-01	0.95 (0.87-1.04)	0.0	7.74e-01
rs3784780	4	genotypic	5.71e-01	1.03 (0.92-1.16)	0.0	8.05e-01

Summary estimates for the 39 SNPs for which data was available in at least two cohorts. Meta-analysis was performed using a fixed effects model in all cases. ES, 95% CI , I^2, and p_Q are included for information only.

Table 4-7 Summary ES and 95%CI for meta-analysis in stage 2 and 3

SNP	Cohorts	Model	p-value	ES (95% CI)	I^2	p_Q
rs7556894	4	recessive	1.33e-06	1.78 (1.41-2.24)	67.3	2.70e-02
rs672757	2	allelic	1.69e-06	1.54 (1.29-1.84)	0.0	3.87e-01
rs472660	3	recessive	7.92e-06	2.54 (1.69-3.83)	69.8	3.66e-02
rs764372	3	recessive	3.96e-05	2.19 (1.51-3.18)	66.1	5.23e-02
rs10429965	3	recessive	1.23e-04	2.60 (1.60-4.23)	78.0	1.06e-02
rs7912136	2	recessive	2.28e-04	1.82 (1.32-2.50)	82.8	1.60e-02
rs437171	2	recessive	2.80e-04	2.28 (1.46-3.56)	76.8	3.79e-02
rs1878632	3	recessive	3.05e-04	1.39 (1.16-1.66)	53.0	1.19e-01
rs1350308	3	allelic	8.96e-04	1.28 (1.11-1.47)	76.7	1.37e-02
rs4649314	4	recessive	1.05e-03	1.29 (1.11-1.51)	67.2	2.76e-02
rs712082	3	recessive	1.12e-03	1.71 (1.24-2.36)	78.8	8.97e-03
rs7007146	3	recessive	1.20e-03	1.34 (1.12-1.60)	85.3	1.12e-03
rs10510044	2	recessive	1.32e-03	1.73 (1.24-2.41)	89.0	2.59e-03
rs1822917	4	recessive	1.75e-03	1.57 (1.18-2.09)	57.1	7.21e-02
rs6972789	3	recessive	2.06e-03	1.32 (1.10-1.57)	89.0	1.12e-04
rs7600624	4	allelic	2.17e-03	1.25 (1.08-1.45)	61.3	5.13e-02
rs567564	3	recessive	2.90e-03	1.43 (1.13-1.82)	84.9	1.33e-03
rs5997921	2	recessive	3.40e-03	2.37 (1.33-4.23)	22.4	2.56e-01
rs1571583	4	recessive	3.71e-03	1.48 (1.14-1.94)	73.9	9.40e-03
rs1438620	2	dominant	4.82e-03	1.45 (1.12-1.88)	83.2	1.47e-02
rs4978394	3	genotypic	6.99e-03	1.22 (1.06-1.41)	70.8	3.26e-02
rs11788150	3	recessive	7.53e-03	1.32 (1.08-1.62)	85.6	9.64e-04
rs1034116	4	allelic	8.57e-03	1.23 (1.05-1.44)	76.6	5.10e-03
rs10842099	2	allelic	8.88e-03	0.76 (0.62-0.93)	90.6	1.13e-03
rs7866165	3	recessive	1.37e-02	1.36 (1.06-1.74)	81.6	4.33e-03
rs4715476	3	genotypic	1.54e-02	1.18 (1.03-1.35)	82.6	3.15e-03
rs4776494	2	genotypic	2.48e-02	1.27 (1.03-1.56)	84.5	1.12e-02
rs9533457	3	recessive	2.90e-02	1.33 (1.03-1.72)	79.6	7.37e-03
rs9315425	3	recessive	2.91e-02	1.69 (1.06-2.71)	73.3	2.36e-02
rs9514816	3	allelic	3.18e-02	1.22 (1.02-1.45)	86.5	5.93e-04
rs3752261	3	recessive	3.18e-02	1.53 (1.04-2.25)	77.7	1.13e-02
rs25689	3	allelic	3.53e-02	1.17 (1.01-1.35)	71.4	3.04e-02
rs1824476	3	allelic	6.00e-02	0.88 (0.76-1.01)	78.8	8.86e-03
rs1924597	4	allelic	6.29e-02	0.85 (0.71-1.01)	84.5	2.34e-04
rs745888	3	recessive	6.60e-02	1.28 (0.98-1.68)	87.0	4.64e-04
rs2514841	3	recessive	1.84e-01	0.88 (0.72-1.07)	78.4	9.66e-03
rs6518956	4	genotypic	2.12e-01	0.94 (0.85-1.04)	0.0	8.21e-01
rs2589183	4	allelic	2.46e-01	0.92 (0.80-1.06)	82.8	5.85e-04
rs3784780	4	genotypic	8.37e-01	0.99 (0.86-1.13)	0.0	7.08e-01

Summary estimates for the 39 SNPs for which data was available in at least two cohorts, confined to stage 2 and 3 patients only. Meta-analysis was performed using a fixed effects model in all cases. ES, 95% CI , I^2, and p_Q are included for information only.

There was high statistical heterogeneity as measured both by I^2 and p_Q, for all but two SNPs. This was largely driven by the VICTOR cohort because the discovery phase had picked SNPs with particularly large ES (the 'winner's curse'), but as the ES in VICTOR

was likely to be an overestimation of the true ES, the heterogeneity measures were not taken into consideration here (see also section 4.3).

4.2.2 Meta-analysis confined to stage 2 and 3 CRC

The SNPs chosen in the discovery phase came from the VICTOR cohort consisting of patients with stage 2 and 3 CRC only. This also was the case for the PETACC cohort, but the Scottish, Danish and Finnish cohorts consisted of patients with all stages of CRC. Therefore, further analysis was undertaken utilising the full data set from both VICTOR and PETACC , but only those Scottish, Danish and Finnish patients with stage 2 and 3 disease to determine if any of the SNPs chosen had a particular effect in this setting.

There were four SNPs (rs7556894, rs672757, rs472660, rs764372) with a p-value lower than the top SNP in the analysis including all stages. The top SNP from the analysis for all stages was no longer the most significant, but was still more significant in this stage restricted analysis (p=7.92e-06). No SNP reached genome-wide significance.

4.2.3 Adjustment for stage

Table 4-8 Meta-analysis for all stages, adjusted for stage

SNP	Cohorts	Model	p-value	ES (95% CI)	I^2	p_Q
rs7556894	4	recessive	1.91e-05	1.60 (1.29-1.98)	75.8	6.14e-03
rs472660	3	recessive	2.63e-05	2.29 (1.56-3.37)	68.4	4.23e-02
rs7912136	2	recessive	3.47e-04	1.67 (1.26-2.22)	86.2	7.00e-03
rs764372	3	recessive	4.97e-04	1.92 (1.33-2.76)	79.0	8.64e-03
rs1878632	3	recessive	5.84e-04	1.33 (1.13-1.57)	66.8	4.91e-02
rs672757	2	genotypic	1.41e-03	1.23 (1.08-1.39)	90.2	1.40e-03
rs7600624	4	genotypic	1.89e-03	1.20 (1.07-1.34)	72.9	1.14e-02
rs4649314	4	recessive	3.50e-03	1.22 (1.07-1.39)	70.2	1.80e-02
rs567564	3	recessive	4.02e-03	1.37 (1.11-1.70)	81.2	4.95e-03
rs1034116	4	genotypic	4.24e-03	1.19 (1.06-1.34)	73.7	9.75e-03
rs10429965	3	recessive	4.45e-03	1.89 (1.22-2.92)	89.8	5.58e-05
rs1571583	4	recessive	5.66e-03	1.40 (1.10-1.78)	70.5	1.72e-02
rs6972789	3	recessive	6.95e-03	1.24 (1.06-1.46)	91.1	1.30e-05
rs1438620	2	dominant	7.04e-03	1.23 (1.06-1.43)	76.8	1.33e-02

Meta-analysis including stage in multivariate analysis, for patients of all stages. Only SNPs with p-value<0.01 are shown. The full Table is given in Appendix A (Table 10-4).

Stage is the strongest known prognostic indicator, and this was therefore included in a multi-variate analysis to test if there is additive prognostic information. Cox regression was performed including stage as a continuous interval variable. As it was not possible to adjust the allelic OR for stage, the stage adjusted genotypic model was used instead. The HR and 95% CI for each cohort are given in Table 4-8 for patients of all stages, and Table 4-9 for stage 2 and 3 disease, the tables are truncated at p≥1.00e-02, the full tables are given in Appendix A.

For the analysis including all stages, most SNPs changed in significance levels (p-value) by less than an order of magnitude, but for eight SNPs this was the case: rs7912136, rs672757, and rs1034116 became more significant; rs4649314, rs437171, rs712082, rs4978394, and rs1924597 became less significant. This change was particularly marked for rs1034116, with a p-value lower by an order of magnitude, and rs1924597 with a p-value higher two orders of magnitude.

For the analysis including only stage 2 and 3, there was very little change in significance and ranking for the SNPs analysed. Only two SNPs changed their p-value by an order of magnitude: rs1350308 and rs437171 both dropped in significance. The 'biggest mover' upwards was rs7600624, although the change in p-value was relatively modest. The significance levels were higher (lower p-values) for all but seven SNPs in the stage 2 and 3 analysis compared to the stage adjusted analysis of all stages.

Table 4-9 Meta-analysis for stage 2 and 3 CRC, adjusted for stage

SNP	Cohorts	Model	p-value	ES (95% CI)	I^2	p_Q
rs7556894	4	recessive	8.96e-07	1.80 (1.42-2.27)	56.9	7.31e-02
rs672757	2	genotypic	1.09e-06	1.47 (1.26-1.72)	38.4	2.02e-01
rs764372	3	recessive	1.25e-05	2.30 (1.58-3.34)	66.4	5.09e-02
rs472660	3	recessive	1.79e-05	2.45 (1.63-3.70)	63.9	6.25e-02
rs10429965	3	recessive	8.52e-05	2.66 (1.63-4.33)	89.9	1.68e-03
rs7912136	2	recessive	1.22e-04	1.87 (1.36-2.56)	80.7	2.28e-02
rs1878632	3	recessive	3.78e-04	1.38 (1.16-1.65)	59.9	8.25e-02
rs7600624	4	genotypic	6.79e-04	1.24 (1.10-1.41)	70.5	1.71e-02
rs4649314	4	recessive	1.34e-03	1.29 (1.10-1.50)	64.4	3.78e-02
rs712082	3	recessive	1.34e-03	1.70 (1.23-2.34)	79.5	7.58e-03
rs6972789	3	recessive	1.59e-03	1.32 (1.11-1.58)	89.9	5.02e-05
rs1571583	4	recessive	1.76e-03	1.53 (1.17-2.00)	67.5	2.65e-02
rs7007146	3	recessive	2.22e-03	1.32 (1.10-1.57)	85.8	8.92e-04
rs5997921	2	recessive	2.59e-03	2.43 (1.36-4.34)	0.0	3.38e-01
rs10510044	2	recessive	4.52e-03	1.61 (1.16-2.24)	78.3	1.00e-02
rs567564	3	recessive	5.37e-03	1.40 (1.10-1.78)	81.8	4.12e-03
rs1822917	4	recessive	7.78e-03	1.47 (1.11-1.96)	54.6	8.53e-02
rs11788150	3	recessive	8.63e-03	1.31 (1.07-1.61)	83.2	2.59e-03
rs1438620	2	dominant	8.96e-03	1.25 (1.06-1.48)	76.6	1.38e-02
rs1350308	3	genotypic	9.22e-03	1.20 (1.05-1.37)	75.8	1.62e-02

Meta-analysis including stage in multivariate analysis, for patients with stage 2 and 3 CRC. Only SNPs with p-value<0.01 are shown. The full Table is given in Appendix A (Table 10-5).

4.3 Meta-analysis for effect size

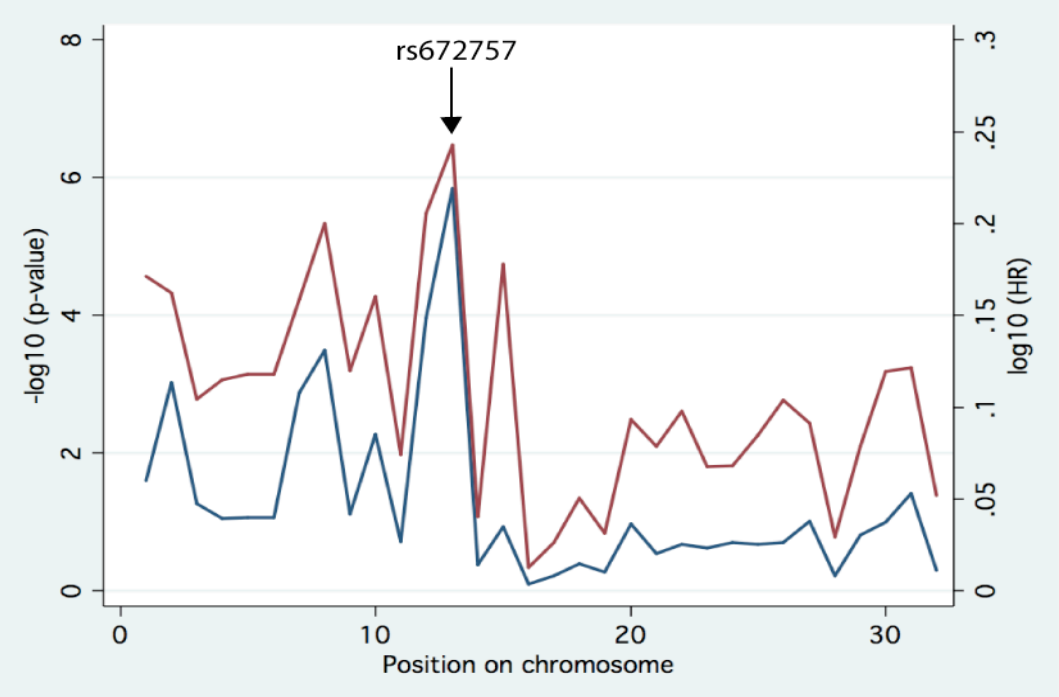
The ES of the loci chosen based on statistical grounds are biased if the same data set is used for locus selection and ES estimation[309]. This is particularly true of GWAS, as the threshold for selection has to be very stringent, and at the same time, the statistical significance is correlated to the ES of the locus in question: for Cox-regression, it is given

by the ratio of the log-transformed ES and its standard error, but other tests of significance are equally affected (Figure 4-2). This effect makes the initial estimate of effect size, but not the p-value, of a particular SNP unreliable[310]. This can lead to the paradoxical phenomenon that a very large effect in the discovery phase reduces the significance levels following the verification phase, the 'winner's curse'. Additionally, empirical data also suggests that the true effect size is likely to be smaller than that first observed[311].

4.3.1 Regression to the mean for the HR and 95% CI of the VICTOR cohort

To adjust for the phenomenon described above, and to derive a more reliable summary estimate of the effect size of the investigated SNPs, a Bayesian analysis of the VICTOR data was performed to derive a more reliable estimate of the true effect size of the SNPs analysed and avoid reporting false positive associations driven by the findings of the VICTOR cohort.

Figure 4-2 Correlation between p-value and HR



Plot of 100Kb either side rs672757 on chromosome 1. p-values are given as -$\log_{10}$ and plotted in navy, HR are given as $\log_{10}$ and plotted in maroon. To aid clarity, the position on chromosome is the number of SNPs typed in this region.

In the absence of a body of evidence for germline determinants of outcome, the prior probability was estimated based on the genetics of CRC risk: there could be a similar number of high-penetrance loci influencing outcome as there are determining CRC risk, about 10 (the equivalent to APC, MMR genes (6 genes), PTEN, LKB1). In addition, there are 13 published SNPs conferring a lesser risk increase of CRC, plus one SNP

described in Chapter 7, and potentially further SNP based loci not yet described. The outside estimate was therefore that there would be no more than 36 loci influencing outcome with a HR=1.2 or greater, and there could be fewer (see also section 2.3.10).

Table 4-10 Adjusted ES for the top 40 SNPs in the VICTOR cohort

SNP	Model	Prior ES (95% CI)	Posterior ES (95% CI)	Posterior p-value
rs472660	recessive	3.95 (2.20-7.09)	1.36 (0.53-3.48)	5.21e-01
rs7556894	recessive	3.03 (2.00-4.59)	2.38 (1.24-4.56)	9.02e-03
rs2589183	allelic	0.48 (0.34-0.67)	0.86 (0.52-1.42)	5.50e-01
rs764372	recessive	3.79 (2.06-6.99)	1.14 (0.61-2.12)	6.88e-01
rs672757	allelic	1.64 (1.31-2.07)	1.11 (0.83-1.50)	4.82e-01
rs7912136	recessive	2.36 (1.61-3.45)	1.26 (0.65-2.44)	4.93e-01
rs6972789	recessive	2.23 (1.65-3.01)	2.00 (1.27-3.13)	2.54e-03
rs7007146	recessive	1.85 (1.37-2.51)	1.08 (0.81-1.45)	6.02e-01
rs745888	recessive	2.35 (1.54-3.58)	1.08 (0.74-1.58)	6.84e-01
rs4978394	genotypic	1.58 (1.24-2.03)	1.05 (0.92-1.21)	4.74e-01
rs3784780	genotypic	0.99 (0.75-1.32)	1.00 (0.93-1.08)	9.86e-01
rs10510044	recessive	2.26 (1.55-3.30)	1.17 (0.68-2.00)	5.75e-01
rs712082	recessive	2.87 (1.80-4.57)	1.31 (0.59-2.91)	5.00e-01
rs11788150	recessive	2.26 (1.59-3.23)	1.34 (0.67-2.67)	4.07e-01
rs10842099	allelic	0.56 (0.42-0.73)	0.91 (0.65-1.26)	5.60e-01
rs4715476	genotypic	1.67 (1.31-2.13)	1.10 (0.83-1.46)	5.11e-01
rs1526884	dominant	0.53 (0.39-0.73)	0.92 (0.68-1.26)	6.03e-01
rs4649314	recessive	2.05 (1.43-2.95)	1.07 (0.79-1.46)	6.58e-01
rs1350308	allelic	1.65 (1.32-2.08)	1.13 (0.82-1.55)	4.72e-01
rs9533457	recessive	2.24 (1.48-3.40)	1.06 (0.79-1.42)	6.99e-01
rs1571583	recessive	2.71 (1.72-4.26)	1.19 (0.63-2.25)	5.94e-01
rs567564	recessive	2.43 (1.67-3.53)	1.53 (0.70-3.34)	2.86e-01
rs7866165	recessive	2.68 (1.67-4.31)	1.10 (0.69-1.75)	6.87e-01
rs437171	recessive	2.99 (1.79-5.00)	1.13 (0.64-1.99)	6.67e-01
rs3752261	recessive	2.90 (1.61-5.21)	1.03 (0.83-1.29)	7.79e-01
rs1924597	allelic	0.36 (0.23-0.56)	0.77 (0.36-1.66)	5.09e-01
rs1822917	recessive	2.39 (1.53-3.74)	1.06 (0.77-1.46)	7.13e-01
rs9514816	allelic	1.96 (1.45-2.65)	1.20 (0.72-2.02)	4.80e-01
rs6518956	genotypic	0.98 (0.80-1.20)	1.00 (0.93-1.07)	9.41e-01
rs1878632	recessive	1.84 (1.33-2.55)	1.05 (0.86-1.28)	6.25e-01
rs7600624	allelic	1.73 (1.29-2.31)	1.05 (0.89-1.24)	5.75e-01
rs9315425	recessive	2.95 (1.56-5.59)	1.02 (0.87-1.21)	7.90e-01
rs10429965	recessive	4.59 (2.49-8.46)	1.75 (0.55-5.53)	3.40e-01
rs5997921	recessive	2.89 (1.48-5.65)	1.02 (0.89-1.16)	7.92e-01
rs1824476	allelic	0.63 (0.48-0.83)	0.96 (0.87-1.07)	4.73e-01
rs4776494	genotypic	1.55 (1.20-2.02)	1.04 (0.94-1.15)	4.51e-01
rs1438620	dominant	2.00 (1.39-2.88)	1.05 (0.83-1.33)	6.68e-01
rs2514841	recessive	0.46 (0.29-0.74)	0.97 (0.84-1.12)	7.14e-01
rs1034116	allelic	1.85 (1.38-2.48)	1.10 (0.79-1.53)	5.81e-01
rs25689	allelic	1.55 (1.20-2.00)	1.04 (0.94-1.16)	4.47e-01

Prior ES and 95% CI derived from latest available survival data (29 April 2009), and the resultant posterior ES and 95% CI after adjustment. The model was retained for each SNP, the p-values were derived using Wald's test. The two SNPs with p<0.05 are highlighted in light grey.

Furthermore, a risk increase of 1.2 would be clinically meaningful, based on the only established prognostic factor, stage, where a risk differential in this order of magnitude is observed between stage 2A and 2B, with the latter generally considered for adjuvant chemotherapy. Based on these assumptions, the adjusted HR and 95% CI were determined for the most significant model for each SNP (see Methods, section 2.3.10).

All SNPs had a lower ES, the SE are smaller for the shrunk estimates than the original estimate for 19 SNPs, while for 21 SNPs the SE of the shrunk estimate is larger, leading to smaller and larger 95% CI, respectively.

The final posterior ES and 95% CI are given in Table 4-10. The adjustment does not allow for a new calculation of Fisher's exact test, used for the allelic model, and the p-value for the allelic model was derived using Wald's test. Following the shrinkage, as expected, none of the SNPs retained their original levels of statistical significance.

4.3.2 Meta-analysis of all stages

Meta-analysis was performed to determine the overall ES for every SNP for which there was data from two or more cohorts. As before, rs1526884 was excluded as this had only been typed successfully in the VICTOR cohort; data for SNPs that deviated from HWE (rs1438620 and rs437171 in PETACC; rs6518956 in the Danish cohort) or had a significantly different genotype distribution from VICTOR (rs10510044 in PETACC) were also not included. The meta-analyses were for the same genetic model as initially detected in the VICTOR cohort, using a fixed effects model. All effect sizes were based on the latest available survival data and are given in Table 4-2, Table 4-3, Table 4-4, Table 4-9, and section 4.1.5.

Two SNPs reached statistical significance at the $p<0.05$ level, rs7556894 with HR=1.32 (95% CI 1.04-1.67, p=2.11e-02) and rs4649314 with HR=1.24 (95%CI 1.04-1.47, p=1.53e-03). Both SNPs had data available from four cohorts, and there was no significant heterogeneity, either by I^2 or p_Q. The summary ES for all SNPs are given in Table 4-10.

There was limited heterogeneity between the adjusted VICTOR data and the verification cohorts, with $I^2>50\%$ for only one SNP. Overall, the suggested effect sizes were relatively modest, with the most extreme effect observed for the SNP that had been the most significant in the initial screening in VICTOR, rs472660 (1.43, 95% CI % CI 0.92-2.25).

Table 4-11 Summary ES and 95%CI for meta-analysis using posterior estimates

SNP	Cohorts	Model	ES (95% CI)	p-value	I^2	p_Q
rs7556894	4	recessive	1.32 (1.04-1.67)	2.02e-02	49.2	1.16e-01
rs4649314	4	recessive	1.16 (1.02-1.32)	2.11e-02	0.0	5.09e-01
rs1924597	4	allelic	0.85 (0.72-1.01)	6.29e-02	0.0	9.76e-01
rs1878632	3	recessive	1.13 (0.98-1.29)	8.77e-02	0.0	5.13e-01
rs472660	3	recessive	1.43 (0.92-2.25)	1.15e-01	42.9	1.73e-01
rs7600624	4	allelic	1.08 (0.97-1.21)	1.64e-01	0.0	9.15e-01
rs4978394	3	genotypic	1.06 (0.96-1.17)	2.65e-01	0.0	9.17e-01
rs1438620	2	dominant	1.10 (0.93-1.30)	2.71e-01	0.0	5.80e-01
rs1824476	3	allelic	0.95 (0.88-1.04)	2.72e-01	44.2	1.67e-01
rs764372	3	recessive	1.20 (0.83-1.74)	3.27e-01	0.0	9.63e-01
rs712082	3	recessive	1.18 (0.85-1.65)	3.29e-01	0.0	9.16e-01
rs6972789	3	recessive	1.08 (0.91-1.28)	3.86e-01	76.5	1.42e-02
rs567564	3	recessive	1.11 (0.87-1.43)	3.98e-01	0.0	6.88e-01
rs1822917	4	recessive	1.09 (0.88-1.36)	4.28e-01	0.0	9.47e-01
rs1571583	4	recessive	1.11 (0.86-1.43)	4.36e-01	0.0	9.13e-01
rs437171	2	recessive	1.16 (0.76-1.77)	4.80e-01	0.0	8.86e-01
rs7866165	3	recessive	1.08 (0.87-1.35)	4.97e-01	0.0	9.52e-01
rs9514816	3	allelic	0.94 (0.78-1.13)	4.99e-01	0.0	5.64e-01
rs4776494	2	genotypic	1.03 (0.94-1.13)	5.23e-01	0.0	6.81e-01
rs2514841	3	recessive	0.97 (0.86-1.09)	6.00e-01	0.0	5.68e-01
rs6518956	4	genotypic	0.99 (0.93-1.05)	6.70e-01	0.0	8.75e-01
rs11788150	3	recessive	1.04 (0.85-1.29)	6.85e-01	0.0	5.93e-01
rs4715476	3	genotypic	1.03 (0.91-1.16)	6.89e-01	0.0	8.52e-01
rs9533457	3	recessive	0.96 (0.78-1.19)	7.05e-01	0.0	3.96e-01
rs10510044	2	recessive	0.94 (0.66-1.33)	7.23e-01	6.0	3.02e-01
rs3784780	4	genotypic	1.01 (0.95-1.08)	7.51e-01	0.0	7.53e-01
rs1034116	4	allelic	0.98 (0.84-1.13)	7.67e-01	0.0	4.95e-01
rs5997921	2	recessive	1.02 (0.90-1.16)	7.72e-01	0.0	9.03e-01
rs25689	3	allelic	1.01 (0.93-1.11)	7.77e-01	27.8	2.50e-01
rs10842099	2	allelic	0.97 (0.81-1.17)	7.83e-01	0.0	6.05e-01
rs10429965	3	recessive	1.06 (0.61-1.83)	8.36e-01	0.0	6.09e-01
rs672757	2	allelic	0.98 (0.83-1.16)	8.45e-01	0.0	3.27e-01
rs2589183	4	allelic	0.99 (0.86-1.13)	8.46e-01	0.0	8.23e-01
rs745888	3	recessive	0.98 (0.78-1.23)	8.74e-01	0.0	8.04e-01
rs7007146	3	recessive	1.01 (0.86-1.19)	8.76e-01	0.0	6.40e-01
rs9315425	3	recessive	1.01 (0.86-1.18)	9.24e-01	0.0	5.11e-01
rs7912136	2	recessive	0.99 (0.69-1.40)	9.34e-01	0.0	3.87e-01
rs1350308	3	allelic	1.00 (0.87-1.15)	9.69e-01	0.0	4.29e-01
rs3752261	3	recessive	1.00 (0.82-1.22)	1.00E+00	0.0	5.32e-01

ES and 95% CI for the 39 SNPs for which there were genotypes available for more than one cohort. Model is the genetic model tested. The rightmost two columns give the measures of heterogeneity, with p_Q referring to the significance of the Q-statistic. Only the data for the VICTOR cohort was adjusted back to the mean.

Table 4-12 Summary ES and 95%CI for meta-analysis in stage 2 and 3 CRC

SNP	Cohorts	Model	ES (95% CI)	p-value	I^2	p_Q
rs7556894	4	recessive	1.52 (1.17-1.96)	1.60e-03	0.0	5.30e-01
rs672757	2	allelic	1.26 (1.03-1.54)	2.79e-02	18.3	2.68e-01
rs472660	3	recessive	1.58 (0.97-2.58)	6.63e-02	21.3	2.81e-01
rs4649314	4	recessive	1.14 (0.99-1.33)	7.65e-02	0.0	6.38e-01
rs764372	3	recessive	1.40 (0.96-2.04)	7.99e-02	0.0	4.47e-01
rs1878632	3	recessive	1.13 (0.98-1.30)	9.90e-02	0.0	5.14e-01
rs7600624	4	allelic	1.09 (0.97-1.23)	1.51e-01	0.0	5.55e-01
rs6972789	3	recessive	1.14 (0.94-1.38)	1.86e-01	73.1	2.44e-02
rs7007146	3	recessive	1.11 (0.93-1.33)	2.27e-01	71.5	3.01e-02
rs1350308	3	allelic	1.09 (0.93-1.28)	2.98e-01	0.0	7.80e-01
rs4978394	3	genotypic	1.06 (0.95-1.18)	3.19e-01	0.0	8.41e-01
rs1822917	4	recessive	1.11 (0.87-1.41)	3.90e-01	0.0	6.84e-01
rs25689	3	allelic	1.04 (0.95-1.13)	4.32e-01	0.0	9.34e-01
rs1034116	4	allelic	1.06 (0.91-1.25)	4.41e-01	0.0	4.37e-01
rs6518956	4	genotypic	0.98 (0.92-1.04)	4.47e-01	0.0	6.01e-01
rs1824476	3	allelic	0.97 (0.89-1.06)	4.67e-01	3.3	3.56e-01
rs1571583	4	recessive	1.11 (0.83-1.48)	5.03e-01	0.0	7.20e-01
rs7866165	3	recessive	1.08 (0.84-1.37)	5.57e-01	0.0	8.64e-01
rs712082	3	recessive	1.11 (0.75-1.64)	5.96e-01	0.0	8.73e-01
rs4776494	2	genotypic	1.03 (0.93-1.13)	5.97e-01	0.0	3.95e-01
rs1438620	2	dominant	1.05 (0.86-1.28)	6.14e-01	0.0	9.94e-01
rs10429965	3	recessive	1.18 (0.61-2.28)	6.31e-01	0.0	7.10e-01
rs7912136	2	recessive	1.11 (0.72-1.71)	6.32e-01	0.0	6.22e-01
rs4715476	3	genotypic	1.03 (0.90-1.19)	6.44e-01	0.0	8.75e-01
rs2589183	4	allelic	1.03 (0.89-1.19)	6.59e-01	0.0	8.63e-01
rs567564	3	recessive	1.06 (0.80-1.41)	6.87e-01	0.0	5.30e-01
rs11788150	3	recessive	1.05 (0.83-1.32)	7.12e-01	0.0	5.41e-01
rs437171	2	recessive	1.09 (0.68-1.76)	7.13e-01	0.0	8.23e-01
rs745888	3	recessive	0.95 (0.74-1.23)	7.18e-01	34.8	2.16e-01
rs1924597	4	allelic	0.97 (0.81-1.17)	7.49e-01	12.0	3.33e-01
rs5997921	2	recessive	1.02 (0.90-1.16)	7.49e-01	0.0	6.40e-01
rs2514841	3	recessive	0.98 (0.87-1.11)	7.80e-01	0.0	7.73e-01
rs10510044	2	recessive	0.95 (0.61-1.45)	7.99e-01	37.9	2.04e-01
rs9514816	3	allelic	0.98 (0.80-1.20)	8.19e-01	0.0	6.83e-01
rs9315425	3	recessive	1.01 (0.86-1.19)	8.68e-01	0.0	5.37e-01
rs3752261	3	recessive	1.02 (0.83-1.25)	8.82e-01	0.0	6.28e-01
rs9533457	3	recessive	1.02 (0.82-1.26)	8.92e-01	0.0	9.13e-01
rs3784780	4	genotypic	1.00 (0.93-1.07)	9.17e-01	0.0	7.02e-01
rs10842099	2	allelic	1.01 (0.81-1.26)	9.37e-01	0.0	3.85e-01

ES and 95% CI for the 39 SNPs for which there were genotypes available for more than one cohort. Model is the genetic model tested. The rightmost two columns give the measures of heterogeneity, with p_Q referring to the significance of the Q-statistic. Only the data for the VICTOR cohort was adjusted back to the mean.

4.3.3 Meta-analysis of stage 2 and 3 disease

As before, a further analysis was undertaken utilising the full data set from both VICTOR and PETACC , but only those Scottish, Danish and Finnish patients with stage 2 and 3 disease. The summary estimates are given in Table 4-12. The significance levels were affected for every SNP, and for 22 SNPs, it decreased, and for 17 SNPs, it increased. Again, there was very little heterogeneity, with I^2>50% for only two SNPs.

Two SNPs, rs7556894 and rs672757, had p-values below the p<0.05 level, while rs4649314, with p<0.05 in the analysis of all stages, no longer was below that level. For rs7556894, the p-value almost reached p<0.05 if it was adjusted for 39 multiple tests using Bonferroni correction.

A multi-variate analysis including stage was also performed in stage 2 and 3 patients for all 39 SNPs, and the HR and 95% CI for SNPs with p<0.1 are given in Table 4-13.

Table 4-13 Meta-analysis for stage 2 and 3, adjusted for stage

SNP	Cohorts	Model	ES (95% CI)	p-value	I^2	p_Q
rs7556894	4	recessive	1.56 (1.21-2.03)	7.25e-04	0.0	6.23e-01
rs672757	2	allelic	1.22 (1.03-1.46)	2.52e-02	0.0	3.46e-01
rs764372	3	recessive	1.44 (0.98-2.09)	6.02e-02	0.0	5.27e-01
rs4649314	4	recessive	1.15 (0.99-1.33)	7.24e-02	0.0	5.08e-01
rs472660	3	recessive	1.53 (0.94-2.49)	9.04e-02	0.0	5.39e-01

Meta-analysis including stage in multivariate analysis, for patients with stage 2 and 3 CRC. Only SNPs with p-value<0.1 are shown. The full Table is given in Appendix A (Table 10-6).

For rs7556894, the summary estimate using all four cohorts (VICTOR, Scotland, PETACC , Denmark) was HR=1.56, 95% CI 1.21-2.03, p=7.25e-04, with no evidence of statistical heterogeneity, I^2=0% and p_Q=0.623. This reaches statistical significance at the Bonferroni corrected p<0.05 level.

4.4 Discussion

The data presented in this chapter support the view that GWAS of germline markers for outcome in CRC are feasible, and that verification cohorts of sufficient quality exist for the verification phase, which, pending further verification, could find novel determinants of prognosis. Given that the VICTOR screening cohort consisted only of stage 2 and 3 patients, the findings for rs7556894 are particularly encouraging, and an effect of other SNPs, e.g. rs672757, rs472660, rs4649314 also cannot be ruled out.

The screening set was included in all meta-analyses as this improves power despite the need for more stringent significance thresholds[274], this is particularly relevant for studies where large numbers of SNPs are taken into the verification phase. While this did not apply to the study design presented in this thesis due to cost restraints, the power of a joint

analysis converges with that of a pure replication analysis with decreasing numbers of SNPs taken into verification, but the former approach is the more powerful one during conversion[274].

For rs7556894, the potential functional implications are intriguing, with genes that may play a role in essential cellular functions for tumorigenesis and metastasis nearby. It is located on the short arm chromosome 2, 15Kb downstream from and in the same LD block as Actin-related protein 2 (ARP2), a member of the ARP2/3 complex essential for cell shape and motility[307] and the process of invasion[312]. rs7556894 was more significant in the stage 2 ad 3 setting, and this is perhaps expected as the initial screen was in these stages. This could be in keeping with the putative functional consequence of any disruption in the tagged gene, i.e. making early metastasis more likely. Patients with stage 2 and 3 CRC homozygous for the minor allele would have a higher chance of being understaged, and micrometastatic distant dissemination would already have taken place at the time of surgery.

No attempt was made as part of this thesis to elucidate the postulated causative loci, nor to validate the functional consequences, and before such work can be undertaken, rs7556894, and other SNPs, will need to be validated further, as none of the p-values reached genome-wide significance levels, generally taken to be $p<1e-07$. This level worked well for the GWAS for CRC susceptibility based on joint analysis, and the SNPs identified at this level were replicated by other groups, most notably rs6983267[38,44,299,313].

While not reaching the necessary significance levels, the fact that the top SNPs became more significant in the analysis restricted to stage 2 and 3 patients, as well as when stage is included in multivariate analysis, suggests that the findings could be real. No attempt was made as part of this thesis to elucidate the postulated causative loci, nor to validate functional consequences.

All meta-analyses for significance were performed using a fixed-effects model, despite the statistical heterogeneity between the cohorts. A random-effects meta-analysis is recommended when the cohorts evaluated do not represent the statistical spectrum of outcomes, while a fixed effects model is appropriate if all cohorts measure the same (single) effect, and if the studies were infinitely large, they would yield the same result[314]. It is likely that the SNPs that are associated with prognosis give a similar effect in each cohort under these assumptions, and the heterogeneity is largely driven by the VICTOR cohort, which, in order to derive highly significant findings, selected very large effect sizes, the so-called winner's curse. The heterogeneity between the verification cohorts was much more modest, in many cases $I^2=0\%$, and only for four SNPs was $I^2>50\%$

(rs1824476, rs4978394, rs1924597, rs9315425), none of which was in the top ten most significant SNPs.

The heterogeneity in terms of effect size and confidence intervals between the original screening cohort and the verification cohorts does make the estimate of the effect size unreliable[310], so that even when the effect was consistently in the same direction as that detected in the VICTOR cohort, but more modest in size, the summary estimate of effect size was likely to be dominated by the VICTOR cohort. The effect size and confidence intervals for the VICTOR cohort were therefore adjusted to derive a better estimate of the potential effect conferred by individual SNPs, likely to be smaller than that detected in VICTOR for statistical[310] and empirical[311] reasons without losing the information included in the VICTOR cohort (an estimate of effect size could have been derived simply by combining the verification cohorts). This particular problem was exacerbated in this GWAS, and likely to be encountered in other GWAS of prognosis using the logrank test in the recessive model, as this allowed for the introduction of very large effect sizes based on relatively few individuals. The estimate of effect size in this setting may be unreliable, and distort the meta-analysis even for those SNPs where there is a true association with prognosis.

The net effect of adjusting the VICTOR effect sizes and confidence intervals was the same as if a further study with ES=1 and an appropriate confidence interval had been added to the meta-analysis, in essence making it a random effects analysis, based on significance levels. In other words, the significance levels of a fixed effects meta-analysis of all available cohorts using the shrunk VICTOR data are similar to those using a random effects meta-analysis with the same cohorts but the unshrunk, prior effects size and CI for VICTOR. In the former, the ES is slightly lower, with slightly smaller CI, making it perhaps a better estimate of the true effect size.

The meta-analysis using shrunk data also gives a separate estimate of the significance levels, but as the effect sizes of the initial screen were adjusted for a possible overestimation, and thus the likelihood of a type 1 error, genome-wide significance at $p<1e-07$ is too conservative. The minimum p-value of the posterior estimates was $p=2.54e-03$, much more in keeping with what would be expected if only 40 SNPs had been analysed, and with what was seen in the Scottish cohort (rs7007146 in genotypic model, $p=4.98e-03$, not analysed here) and PETACC (rs472660 in recessive model, $p=2.15e-02$). A Bonferroni corrected $p<0.05$ for 39 SNPs would be more appropriate to test significance in the shrunk analysis, equivalent to $p<1.28e-03$ in the shrunk analysis, which rs7556894 reached only if also adjusted for stage.

The low rate of replication was to be expected as only 14 SNPs had an ES going in the same direction in all available cohorts, and in only seven of these (rs764372, rs712082, rs4649314, rs567564, rs1924597, rs1878632, rs7600624) was the ES not essentially 1 in one or the other of the verification cohorts. Only two SNPs had an ES of less than 0.85 (or greater than 1.15) in all three cohorts: rs4649314 and rs764372. The latter did demonstrate that the shrinkage yielded reasonable estimates, as the posterior estimate for this SNP in the VICTOR cohort was HR=1.14 (95% CI 0.61-2.12, p=6.87e-01), very similar to the Scottish cohort HR=1.19 (95% CI 0.61-2.31, p=6.08e-01) and PETACC , HR=1.28 (95% CI 0.69-2.41, p=4.34e-01).

In all but two SNPs, rs6518956 and rs3784780, the statistical heterogeneity introduced by using unshrunk priors from the VICTOR cohort yielded I^2>50%. Both these SNPs were initially included based on the highly significant logrank test in the genotypic model, but with much higher p-values when analysed by Cox regression. This was due to the worsening survival from *Mm* to *mm* to *MM* genotypes, thus the coding was not 'in order' of the number of minor alleles. This effect was seen both in VICTOR and the Scottish cohort for rs3784780, but not in PETACC , while it was only seen in VICTOR for rs6518956. It suggests that the failure to replicate the association seen in VICTOR for both these SNPs was because of a false positive finding, rather than an interesting effect on tumour biology not capture by the coding of alleles and subsequent Cox regression.

Some of the problems that arose from the initial selection strategy have been addressed by using shrunk posterior estimates for VICTOR, and it is likely that the final estimates of significance and effect size are reasonable, although further verification should be attempted for the top SNPs, rs7556894, at a minimum. However, it is likely that too many SNPs were being taken into the verification phase without sufficient evidence to support them, a result of allowing several models to be considered, and relatively low event rates. Therefore, an approach that restricts the initial SNP selection to Cox regression based on the genotypic model, includes stage in the initial discovery phase, and using larger cohorts with higher event rates might yield more germline SNPs that in the verification phase remain associated with prognosis. The discovery phase for such an approach is presented in Chapter 5. Based on the data presented in Chapter 3 and this chapter, an effect of the germline on prognosis in CRC remains possible, but is not proven.

Chapter 5

Alternative strategies for the discovery phase

Different screening or discovery strategies can be derived, either to try to improve the discovery of true-positives in the search for prognostic determinants, or to try to discover predictive determinants. Based on the findings in Chapters 3 and 4, two such approaches are described in this chapter.

5.1 Meta-analysis of two screening sets

Given the limitations of the approach described in Chapters 3 and 4, a different screening approach was conceived to address some of the issues encountered and at the same time, increase the power for detection. Genotypes from the Scottish cohort utilised in Chapter 4 were available for almost all SNPs analysed in the initial screen in the VICTOR cohort (see Chapter 3). Therefore, it was possible to meta-analyse the VICTOR and Scottish cohorts for all SNPs to increase sample size and event rate, and trying to address some of the limitations encountered so far.

5.1.1 Analysis of individual cohorts

Each autosomal SNP was analysed by Cox regression for the genotypic model only to avoid the over-representation of recessive model SNPs in the final analysis. The initial survival analysis for each cohort was adjusted for stage, age and sex to avoid detecting SNPs that might be associated with stage rather than prognosis *per se*.

In CRC, the HR generated from 3-year DFS analysis compared to 5-year OS analysis are tightly correlated (r^2=0.90)[315], and, on the assumption of proportionality within the Cox model, should be at all time points. Increasing the event number should drive more precise estimates, and therefore DFS was used as outcome measure for VICTOR, while CRC specific survival was used for the Scottish cohort.

The survival analysis was conducted separately in each cohort, and the outcome measures meta-analysed using a fixed effects model. Significant SNPs at a level p<1e-04 displaying significant heterogeneity, defined as I^2>50%, were further analysed using a random effects model.

5.1.1.1 VICTOR

The data cut-off for the Victor cohort was 29 April 2009, and 947 patients were included in the survival analysis. This cohort and outcome data are identical to that described in section 4.1.3.

Table 5-1 Top SNPs in VICTOR cohort

SNP	Chr	Position	HR (95% CI)	p-value	Function
rs2144749	6	168.70	1.64 (1.34-2.02)	2.52e-06	SPARC related calcium binding 2
rs5761880	21	25.69	2.08 (1.53-2.82)	3.04e-06	cDNA FLJ31418 (80kb)
rs9514816	13	107.58	0.55 (0.42-0.71)	5.26e-06	DNA ligase IV
rs672757	1	112.15	0.62 (0.50-0.76)	5.76e-06	K+- voltage-gated channel
rs2827250	21	22.43	0.55 (0.42-0.71)	6.04e-06	clone qd65g07 PRED16 (17kb)
rs17808334	18	66.02	0.36 (0.23-0.56)	6.54e-06	Rotatin
rs4891821	18	66.00	2.72 (1.75-4.24)	9.70e-06	Rotatin
rs1924597	13	96.24	0.39 (0.25-0.60)	1.78e-05	HS6ST3
rs806711	19	5.57	0.51 (0.37-0.70)	2.30e-05	Scaffold attachment factor B2
rs12275549	11	94.68	0.53 (0.40-0.71)	2.49e-05	AW966820
rs2589183	8	97.59	1.93 (1.42-2.62)	2.86e-05	Syndecan 2 precursor
rs3124028	9	7.66	1.56 (1.27-1.93)	3.34e-05	cDNA FLJ46908 (130kb)
rs6491307	13	96.25	1.98 (1.43-2.73)	3.39e-05	HS6ST3
rs7864533	9	29.58	0.60 (0.47-0.77)	4.18e-05	Nothing within 750kb
rs806702	19	5.58	0.52 (0.38-0.71)	4.20e-05	Scaffold attachment factor B2
rs4715476	6	54.73	1.65 (1.30-2.10)	4.67e-05	Tubulointerstitial nephritis antigen(370kb)
rs3741355	11	3.08	0.55 (0.42-0.74)	4.72e-05	Oxysterol-binding protein like 5
rs1041795	21	38.77	1.58 (1.27-1.97)	4.77e-05	Ets-related isoform 1
rs6972789	7	66.76	1.55 (1.25-1.92)	5.67e-05	Stromal antigen 3-like 4 (350kb)
rs11120735	1	214.35	0.57 (0.43-0.75)	5.79e-05	Usherin isoform B
rs958882	5	123.03	1.56 (1.25-1.93)	5.83e-05	Casein kinase 1, γ3 isoform 4 (47kb)
rs17677737	7	29.15	2.04 (1.44-2.90)	6.36e-05	Serine carboxypeptidase vitellogenic-like
rs7798099	7	141.77	0.57 (0.44-0.75)	6.68e-05	mRNA for T cell receptor beta chain
rs7600624	2	152.99	1.66 (1.29-2.14)	7.33e-05	Formin-like 2
rs7229842	18	65.89	2.18 (1.48-3.20)	7.60e-05	Rotatin
rs4239387	18	29.92	0.61 (0.47-0.78)	7.65e-05	Nucleolar protein 4
rs10788500	10	88.15	0.59 (0.46-0.77)	7.75e-05	Wings apart-like homolog (40kb)
rs10281994	7	125.27	1.74 (1.32-2.29)	8.17e-05	Glutamate receptor isoform c (550kb)
rs6949266	7	125.18	1.76 (1.33-2.33)	8.69e-05	Glutamate receptor isoform c (550kb)
rs10451282	17	41.14	0.43 (0.28-0.66)	9.59e-05	CRH receptor 1 (75kb)

Thirty SNPs meeting significance threshold of p≤1e-04 by Cox regression in genotypic model, adjusted for stage, age and sex. The 30 SNPs represent 25 independent loci. The 7 SNPs selected in the initial screen are highlighted in light grey. Chr - chromosome.

Thirty SNPs had p<1e-04, they cover 25 independent loci. Thirteen loci are within genes, while 12 are in intergenic regions (Table 5.1). Only seven SNPs had been selected by the initial screen described in Chapter 3.

While the most significant SNPs in this analysis were not being treated differently from any other SNP, there was again a number of candidate SNPs in interesting areas. In addition to the SNPs already described in Chapter 3, rs2144749, the most significant SNP in this analysis, is located in SPARC related modular calcium binding 2, a widely-expressed matricellular protein which contributes to mitogenesis via activation of Integrin-Linked Kinase[316]. rs12275549 is located 75kb from sestrin 3, a TP53 responsive anti-oxidant that decreases the mutagenic effects of RAS induced intracellular reactive oxygen species[317], while at the same time methylation of the CpG islands of sestrin 3 is strongly associated with *MLH1* methylation[318].

5.1.1.2 SCOTTISH COHORT

Forty-one SNPs had p<1e-04, they cover 34 independent loci. 10 loci are within genes, while 24 are in intergenic regions (Table 5.2). There was no overlap between the SNPs identified in this screen and either of the screens performed in the VICTOR cohort (section 5.1.1 and Chapter 3).

Like for the VICTOR cohort, there were several interesting loci among the most significant SNPs. The most significant SNP, rs4074853, is located in PTPN5, a non-receptor type protein tyrosine phosphatase, which inactivates MAPK[319]. rs3136551 lies just upstream of CD27, a member of the TNF-receptor superfamily, which upon binding of its ligand CD70 leads to the activation of NF-κB[320]. rs2413485 is located only 7kb from histone 1, family member 0 (*H1F0*), and tumour derived factors may inhibit dendritic cell differentiation by affecting *H1F0* expression, thus decreasing the immune response to tumour related antigen[321].

Again, all SNPs from this analysis were treated the same and meta-analysed with the results of the VICTOR cohort.

5.1.2 Meta-analysis

Data from both cohorts were available for 300,370 SNPs. 150,612 SNPs had an HR in the same direction in both cohorts, making these the likely candidates for significant findings. Of these, 564 SNPs had significant heterogeneity at the uncorrected $p_{het}<0.05$, and a further 3760 SNPs had $I^2 \geq 50\%$.

Thirty-four SNPs met the significance threshold of p<1e-04, out of these, all but one SNP were in HWE in both cohorts. rs9810699 was not in HWE in the overall and control cohorts in VICTOR and was excluded. The 33 SNPs cover 30 independent loci. 17 loci

Table 5-2 Top SNPs in Scottish cohort

SNP	Chr	Position	HR (95% CI)	p-value	Nearest gene
rs4074853	11	18.749	0.44 (0.33-0.59)	3.04e-08	PTPN5, non-receptor type
rs10027059	4	181.356	0.58 (0.46-0.73)	5.09e-06	Nothing within 750kb
rs2280400	17	35.603	0.58 (0.46-0.74)	6.49e-06	Rap guanine nucleotide exchange factor
rs10500624	11	4.905	1.53 (1.27-1.86)	1.07e-05	Olfactory receptor, F51, subfamily G (2kb)
rs4861488	4	182.821	0.70 (0.59-0.82)	1.53e-05	cDNA FLJ31634 (204kb)
rs4812851	20	42.754	1.44 (1.22-1.71)	2.27e-05	cDNA FLJ33523
rs6906061	6	85.198	0.69 (0.58-0.82)	2.41e-05	KIAA1009 protein (104kb),
rs895796	1	30.141	0.58 (0.45-0.74)	2.42e-05	cDNA clone IMAGE:4830073 (118kb)
rs6803929	3	43.348	1.84 (1.38-2.45)	2.66e-05	SNF related kinase
rs6788594	3	194.972	0.59 (0.46-0.76)	2.95e-05	optic atrophy 1 isoform 8 (73kb)
rs4979603	9	117.778	2.11 (1.48-2.99)	3.20e-05	EST-YD1 mRNA (50kb)
rs7741966	6	85.239	0.69 (0.58-0.83)	3.56e-05	KIAA1009 protein (104kb)
rs1860339	17	58.205	0.72 (0.61-0.84)	3.80e-05	Ring finger protein 190
rs2570583	11	4.921	1.51 (1.24-1.84)	3.93e-05	Olfactory receptor, F51, subfamily G (2kb)
rs3136551	12	6.424	2.36 (1.56-3.55)	4.01e-05	TNF receptor superfamily (0.2kb)
rs9449923	6	85.222	1.43 (1.20-1.70)	5.10e-05	KIAA1009 protein (104kb)
rs6676127	1	33.805	1.42 (1.20-1.68)	5.16e-05	CUB and Sushi multiple domains 2
rs2688492	3	196.982	1.44 (1.21-1.71)	5.25e-05	MUC4
rs4812456	20	38.840	0.69 (0.58-0.83)	5.39e-05	EST DB112410
rs1550790	16	58.442	0.69 (0.57-0.83)	5.63e-05	Nothing within 750kb
rs1534995	6	41.069	0.68 (0.56-0.82)	6.26e-05	UNC5CL (33kb)
rs9374571	6	116.207	0.70 (0.59-0.83)	6.26e-05	Fyn-related kinase (162kb)
rs1445458	16	58.436	0.68 (0.56-0.82)	6.37e-05	Nothing within 750kb
rs4394	22	47.055	0.62 (0.49-0.78)	6.51e-05	FAM19A5 (208kb)
rs1572372	13	19.738	0.70 (0.59-0.84)	6.70e-05	Gap junction protein, beta 6 (44kb)
rs2902374	2	82.703	0.65 (0.53-0.80)	7.13e-05	Nothing within 750kb
rs2413485	22	36.524	1.41 (1.19-1.66)	7.56e-05	H1 histone family, member 0 (7kb)
rs4381631	17	62.197	0.72 (0.61-0.84)	7.58e-05	Protein kinase C, alpha variant
rs6818678	4	157.417	0.67 (0.55-0.82)	7.67e-05	PDGF C precursor (485kb)
rs1046726	5	178.248	0.71 (0.60-0.84)	7.83e-05	Zinc finger protein 354B (29kb)
rs3782181	12	87.478	1.44 (1.20-1.73)	7.98e-05	KIT ligand isoform b precursor
rs6507839	18	44.091	0.67 (0.55-0.82)	8.09e-05	zinc finger and BTB domain containing 7C
rs4694656	4	75.030	1.58 (1.26-1.99)	8.27e-05	platelet factor 4 (35kb)
rs4438512	2	71.499	1.39 (1.18-1.64)	8.67e-05	zinc finger protein 638
rs7562587	2	71.516	1.39 (1.18-1.64)	8.67e-05	zinc finger protein 638
rs2825929	21	20.278	0.72 (0.61-0.85)	8.67e-05	nothing
rs2407581	21	20.273	0.72 (0.61-0.85)	8.75e-05	nothing
rs4946118	6	116.204	0.71 (0.59-0.84)	9.02e-05	Fyn-related kinase (162kb)
rs2304218	19	44.651	0.45 (0.30-0.67)	9.22e-05	Transcription elongation factor SPT5
rs6431734	2	15.792	0.69 (0.57-0.83)	9.37e-05	cDNA FLJ36206 (21kb)
rs987189	17	14.492	0.58 (0.44-0.76)	9.95e-05	DNA FLJ45831 (119kb)

Forty-one SNPs with p≤1e-04 by Cox regression in genotypic model, adjusted for stage, age and sex. The 41 SNPs represent 34 independent loci. Chr - chromosome. The nearest gene is either the gene within which the SNP is located, or the nearest transcript to it.

Table 5-3 Top SNPs from the meta-analysis of VICTOR and Scottish cohort

SNP	Chr	Position	HR (95% CI)	p-value	I^2	$HR_{Scotland}$ (95% CI)	$p_{Scotland}$	HR_{VICTOR} (95% CI)	p_{VICTOR}	Function
rs4074853	11	18.749	0.54 (0.43-0.68)	2.11e-07	80.7	0.44 (0.33-0.59)	3.04e-08	0.78 (0.53-1.17)	2.32e-01	Protein tyrosine phosphatase, non-receptor type
rs2280400	17	35.603	0.64 (0.54-0.77)	1.80e-06	42.5	0.58 (0.46-0.74)	6.49e-06	0.74 (0.56-0.99)	4.03e-02	Rap guanine nucleotide exchange factor
rs10500624	11	4.905	1.46 (1.25-1.70)	2.44e-06	0.0	1.53 (1.27-1.86)	1.07e-05	1.31 (1.00-1.72)	5.40e-02	Olfactory receptor, 51G (2kb)
rs2570583	11	4.921	1.45 (1.24-1.70)	4.33e-06	0.0	1.51 (1.24-1.84)	3.93e-05	1.35 (1.03-1.77)	3.12e-02	Olfactory receptor, 51G (2kb)
rs240486	21	16.694	0.68 (0.57-0.80)	4.89e-06	0.0	0.65 (0.52-0.81)	1.27e-04	0.72 (0.55-0.92)	1.05e-02	LOC388815 isoform a
rs2837132	21	40.026	1.51 (1.26-1.80)	5.55e-06	6.8	1.41 (1.13-1.75)	2.45e-03	1.71 (1.27-2.30)	4.00e-04	Immunoglobulin superfamily 5 like (13kb)
rs729552	20	12.131	0.71 (0.61-0.83)	6.58e-06	0.0	0.75 (0.61-0.91)	3.67e-03	0.67 (0.54-0.84)	4.27e-04	BTB/POZ domain containing protein 3 (312kb)
rs10487167	7	103.163	1.77 (1.37-2.28)	1.36e-05	0.0	1.87 (1.34-2.62)	2.55e-04	1.63 (1.10-2.42)	1.59e-02	Reelin isoform a
rs672740	1	170.404	0.76 (0.67-0.86)	1.63e-05	0.0	0.74 (0.63-0.87)	2.42e-04	0.79 (0.65-0.96)	2.02e-02	Dynamin 3 isoform a
rs4381631	17	62.197	0.76 (0.66-0.86)	2.09e-05	2.1	0.72 (0.61-0.84)	7.58e-05	0.82 (0.67-1.01)	6.27e-02	Protein kinase C, alpha variant
rs4861488	4	182.821	0.76 (0.66-0.86)	2.25e-05	59.8	0.70 (0.59-0.82)	1.53e-05	0.87 (0.70-1.07)	1.85e-01	cDNA FLJ31634 (204kb)
rs895796	1	30.141	0.67 (0.55-0.80)	2.26e-05	62.1	0.58 (0.45-0.74)	2.42e-05	0.79 (0.60-1.04)	9.59e-02	cDNA clone IMAGE:4830073 (118kb)
rs1004361	12	116.476	0.68 (0.57-0.81)	2.35e-05	0.0	0.70 (0.56-0.89)	3.00e-03	0.65 (0.49-0.86)	2.35e-03	Kinase suppressor of ras 2
rs10027059	4	181.356	0.65 (0.53-0.80)	2.60e-05	71.8	0.58 (0.46-0.73)	5.09e-06	0.88 (0.61-1.28)	5.12e-01	Nothing within 750kb
rs7600624	2	152.991	1.43 (1.21-1.70)	3.24e-05	59.4	1.27 (1.01-1.59)	4.52e-02	1.66 (1.29-2.14)	7.33e-05	Formin-like 2
rs2640882	6	116.882	1.33 (1.16-1.52)	3.25e-05	0.0	1.35 (1.14-1.60)	5.17e-04	1.29 (1.04-1.61)	2.12e-02	Dermatan sulfate epimerase precursor
rs4812456	20	38.840	0.74 (0.65-0.86)	3.33e-05	40.3	0.69 (0.58-0.83)	5.39e-05	0.83 (0.67-1.04)	1.08e-01	EST DB112410
rs6853554	4	107.167	1.38 (1.18-1.61)	3.48e-05	57.4	1.25 (1.03-1.52)	2.25e-02	1.60 (1.25-2.04)	1.58e-04	TBC domain-containing kinase-like (19kb)
rs1396087	17	49.668	1.38 (1.18-1.61)	3.97e-05	0.0	1.45 (1.19-1.77)	2.13e-04	1.27 (1.00-1.63)	5.01e-02	Kinesin family member 2B (411kb)
rs2385168	7	103.866	0.74 (0.65-0.86)	3.97e-05	0.0	0.71 (0.59-0.85)	1.48e-04	0.81 (0.64-1.02)	6.76e-02	Lipoma HMGIC fusion partner-like 3
rs3782176	12	87.463	0.68 (0.57-0.82)	4.29e-05	0.0	0.63 (0.50-0.80)	1.67e-04	0.76 (0.57-1.02)	6.36e-02	KIT ligand isoform b precursor
rs1828096	17	49.668	0.73 (0.62-0.85)	4.52e-05	0.0	0.69 (0.56-0.84)	1.96e-04	0.79 (0.62-1.01)	5.98e-02	Kinesin family member 2B (411kb)

Table 5-3 Top SNPs from the meta-analysis of VICTOR and Scottish cohort (cont'd)

SNP	Chr	Position	HR (95% CI)	p-value	I^2	$HR_{Scotland}$ (95% CI)	$p_{Scotland}$	HR_{VICTOR} (95% CI)	p_{VICTOR}	Function
rs6803929	3	43.348	1.61 (1.28-2.03)	5.13e-05	59.0	1.84 (1.38-2.45)	2.66e-05	1.25 (0.84-1.86)	2.75e-01	SNF related kinase
rs1326450	13	104.330	0.76 (0.67-0.87)	5.91e-05	0.0	0.76 (0.64-0.90)	1.39e-03	0.77 (0.63-0.95)	1.49e-02	D-amino acid oxidase activator ()585kb
rs4779800	15	27.199	1.47 (1.22-1.77)	6.18e-05	0.0	1.49 (1.16-1.91)	1.63e-03	1.43 (1.08-1.90)	1.31e-02	Amyloid β A4 precursor protein-binding (2kb)
rs2296441	10	100.135	1.34 (1.16-1.55)	6.87e-05	0.0	1.34 (1.12-1.61)	1.38e-03	1.34 (1.05-1.71)	1.79e-02	Pyridine nucleotide-disulphide oxidoreductase domain 2
rs966534	20	5.736	0.74 (0.63-0.86)	7.02e-05	0.0	0.75 (0.62-0.92)	5.05e-03	0.71 (0.56-0.90)	4.53e-03	Hypothetical protein LOC149840
rs6849573	4	107.243	1.36 (1.17-1.58)	7.11e-05	39.4	1.25 (1.03-1.52)	2.24e-02	1.54 (1.21-1.96)	4.73e-04	TBC domain-containing protein kinase-like
rs10883450	10	101.913	1.69 (1.30-2.18)	7.71e-05	31.5	1.46 (1.03-2.07)	3.56e-02	2.01 (1.37-2.94)	3.72e-04	ER lipid raft associated 1
rs988421	1	72.322	0.77 (0.68-0.88)	8.25e-05	0.0	0.75 (0.64-0.89)	7.77e-04	0.80 (0.65-0.99)	3.58e-02	Neuronal growth regulator 1
rs7424907	2	69.158	1.49 (1.22-1.82)	8.82e-05	0.0	1.61 (1.25-2.07)	2.34e-04	1.31 (0.95-1.81)	9.60e-02	Anthrax toxin receptor 1 isoform 3 precursor
rs7752055	6	106.067	1.49 (1.22-1.81)	9.02e-05	0.0	1.53 (1.18-1.99)	1.45e-03	1.43 (1.05-1.94)	2.13e-02	Prolyl endopeptidase (70kb)
rs10459149	12	90.937	0.71 (0.60-0.85)	9.90e-05	0.0	0.68 (0.55-0.86)	1.07e-03	0.75 (0.58-0.97)	2.98e-02	Full length insert cDNA clone YW25A12

Thirty-three SNPs meeting significance threshold of p≤1e-04 by fixed effects meta-analysis. These SNPs represent 30 independent loci. The 10 SNPs picked in the most significant SNPs from the Scottish cohort are highlighted in light grey, the single SNP from the VICTOR cohort in darker grey.

were within genes, while 13 were in intergenic regions. Out of the top 33 SNPs, one had been present in the top SNPs from the VICTOR screen (rs7600624), and 10 in the top SNPs from the Scottish cohort screen (rs4074853, rs2280400, rs10500624, rs2570583, rs4381631, rs4861488, rs895796, rs10027059, rs4812456, rs6803929); see Table 5-3.

Seven SNPs had $I^2 \geq 50\%$ and were also analysed by a random effects model (rs4074853 rs4861488 rs895796 rs10027059 rs7600624 rs6853554 rs6803929). In all cases, heterogeneity was driven by the HR from one cohort not being within the 95% CI of the other cohort. As expected, for all SNPs, in random effects meta-analysis, the 95% CI became wider and the significance levels decreased, with higher p-values, while the summary HR only change changed by a small degree (Table 5-4).

Table 5-4 Random effects meta-analysis for SNPs with $I^2 \geq 50\%$

SNP	HR (95% CI)	p-value	I^2	$HR_{Scotland}$ (95% CI)	$p_{Scotland}$	HR_{VICTOR} (95% CI)	p_{VICTOR}
rs4074853	0.58 (0.33-1.01)	5.50e-02	80.7	0.44 (0.33-0.59)	3.04e-08	0.78 (0.53-1.17)	2.32e-01
rs4861488	0.77 (0.62-0.95)	1.40e-02	59.8	0.70 (0.59-0.82)	1.53e-05	0.87 (0.70-1.07)	1.85e-01
rs895796	0.67 (0.49-0.91)	1.10e-02	62.1	0.58 (0.45-0.74)	2.42e-05	0.79 (0.60-1.04)	9.59e-02
rs10027059	0.70 (0.46-1.05)	8.50e-02	71.8	0.58 (0.46-0.73)	5.09e-06	0.88 (0.61-1.28)	5.12e-01
rs7600624	1.44 (1.11-1.89)	7.15e-03	59.4	1.27 (1.01-1.59)	4.52e-02	1.66 (1.29-2.14)	7.33e-05
rs6853554	1.40 (1.10-1.78)	5.44e-03	57.4	1.25 (1.03-1.52)	2.25e-02	1.60 (1.25-2.04)	1.58e-04
rs6803929	1.56 (1.07-2.27)	2.20e-02	59.0	1.84 (1.38-2.45)	2.66e-05	1.25 (0.84-1.86)	2.75e-01

The heterogeneity is largely driven by higher significance levels in the Scottish cohort, although for two SNPs, this is the case for VICTOR, including rs7600624, which had been selected in the initial discovery phase presented in Chapter 3.

The top SNP in the meta analysis was the same as that from the Scottish cohort, rs4074853 located in *PTPN5*, but this also had the highest degree of statistical heterogeneity, making this SNP not significant a random effects analysis. Other SNPs are of biological interest, but given the low replication rate of those SNPs chosen on biological rationale in Chapters 3 and 4, this should not be a criterion for selection of SNPs for further verification. This verification has not yet been not performed, but further suitable cohorts are actively being identified; it is therefore not presented in this thesis.

5.2 Screening for loci predictive of benefit from 5-FU chemotherapy

As well as screening for prognostic markers, the samples of the VICTOR cohort also provided an opportunity for the discovery of predictive markers. Therefore, the subset of the VICTOR cohort that had received adjuvant chemotherapy was analysed separately for loci that might carry such predictive information.

5.2.1 GWAS for predictive markers

Of the total VICTOR cohort, 595 patients received adjuvant 5-FU based chemotherapy, 151 with stage 2 and 444 with stage 3 CRC. At the data cut-off point on 29 April 2009, there were 135 relapses after a median follow-up of 60.2 months (SD ±12.2 months).

Table 5-5 Top predictive SNPs in VICTOR

SNP	Chr	Position	HR (95% CI)	p-value	Nearest gene
rs4715476	6	54730243	1.89 (1.61 - 2.17)	7.17e-06	Tubulointerstitial nephritis antigen (370kb)
rs6117279	20	646986	1.72 (1.48 - 1.97)	1.04e-05	Sulfiredoxin 1 homolog (60kb)
rs1408811	9	12565420	2.22 (1.86 - 2.58)	1.39e-05	Tyrosinase-related protein 1(120kb)
rs917205	7	87555697	2.32 (1.94 - 2.71)	1.52e-05	ADAM metallopeptidase
rs1516708	4	92721389	0.57 (0.31 - 0.83)	1.87e-05	*FAM190A*, transcript variant 1
rs968757	4	120976901	1.69 (1.45 - 1.93)	2.02e-05	Phosphodiesterase 5A isoform 2 (209kb)
rs7924634	11	132932946	1.91 (1.61 - 2.21)	2.27e-05	Opioid binding /cell adhesion (25kb)
rs5743291	16	49314776	2.17 (1.81 - 2.52)	2.29e-05	Nucleotide-binding oligomeriszation
rs7787525	7	99331595	2.42 (2.01 - 2.83)	2.33e-05	*TRIM4* isoform
rs510902	2	238992521	1.68 (1.44 - 1.92)	2.58e-05	Ankyrin repeat/SOCS box (8kb)
rs746017	7	130955479	0.51 (0.19 - 0.82)	2.61e-05	Podocalyxin-like isoform precursor (64kb)
rs3124238	9	89436136	1.66 (1.42 - 1.89)	3.05e-05	Death-associated protein kinase 1
rs10281994	7	125272031	1.90 (1.60 - 2.21)	3.19e-05	Glutamate receptor, isoform c (590kb)
rs2002059	10	101120844	2.32 (1.92 - 2.71)	3.53e-05	Cyclin M1
rs2144749	6	168695539	1.67 (1.43 - 1.91)	3.61e-05	SPARC related modular calcium binding 2
rs3784929	16	74234527	1.82 (1.54 - 2.11)	3.77e-05	Lysyl-tRNA synthetase isoform 3
rs10102339	8	27023404	1.66 (1.42 - 1.91)	3.89e-05	Stathmin-like 4 (126kb)
rs13253769	8	100102019	0.59 (0.34 - 0.84)	4.08e-05	Vacuolar protein sorting 13B isoform 4
rs1041795	21	38774265	1.70 (1.44 - 1.95)	4.38e-05	Ets-related isoform 1
rs8053257	16	74232188	2.03 (1.69 - 2.37)	4.94e-05	Lysyl-tRNA synthetase isoform 2
rs883834	2	45310614	2.07 (1.72 - 2.43)	5.01e-05	Homo sapiens cDNA FLJ38231
rs9514816	13	107581122	1.89 (1.58 - 2.20)	5.78e-05	DNA ligase IV (75kb)
rs4923791	15	36476362	1.97 (1.64 - 2.31)	6.41e-05	Sprouty-related protein 1 (90kb)
rs10913063	1	173942218	1.98 (1.64 - 2.31)	6.46e-05	Tenascin R
rs237176	16	26567930	0.55 (0.26 - 0.84)	6.79e-05	Heparan sulfate D-glucosaminyl (510kb)
rs7591253	2	34840709	1.67 (1.42 - 1.93)	7.06e-05	cDNA clone CS0DI009YL06 (39kb)
rs2013746	2	7328172	1.79 (1.50 - 2.08)	7.35e-05	Ring finger protein 144 (227kb)
rs337532	9	100490229	2.21 (1.82 - 2.61)	7.52e-05	G protein-coupled receptor 51
rs1819887	18	50654310	1.67 (1.42 - 1.93)	7.90e-05	Member RAS oncogene family 27B
rs6491307	13	96252810	0.47 (0.09 - 0.85)	8.10e-05	Heparan sulfate 6-O-sulfotransferase 3
rs10934005	3	111185514	1.71 (1.44 - 1.97)	8.42e-05	Development pluripotency assoc'd (645kb)
rs5764560	22	42876697	0.57 (0.29 - 0.85)	8.53e-05	Parvin, beta isoform a
rs10161354	12	87869619	2.25 (1.84 - 2.65)	8.73e-05	Homo sapiens cDNA FLJ30500 (60kb)
rs12419726	11	59495242	1.80 (1.50 - 2.09)	9.17e-05	mRNA for Hsa11-digit01-13-09-F
rs17596719	6	26205172	1.93 (1.60 - 2.26)	9.49e-05	Haemochromatosis splice variant (0.1kb)
rs1892431	1	90143395	1.81 (1.51 - 2.11)	9.61e-05	Leucine rich repeat containing 8 family
rs1730265	18	49843033	1.88 (1.56 - 2.20)	9.66e-05	Methyl-CpG binding protein 2 (90kb)
rs11709756	3	103334906	1.89 (1.57 - 2.21)	9.90e-05	Zona pellucida-like domain containing 1

SNPs representing 38 independent loci. The nearest gene is either the gene within which the SNP is located, or the nearest transcript to it. The genetic model is genotypic for all SNPs. Chr - chromosome.

Unlike the analysis presented in section 5.1, there was no dataset available that had information on chemotherapy and genotypes for all SNPs on the Hap300 arrays. SNPs with a minor allele frequency of less than 5% were excluded to avoid the difficulties encountered in the discovery phase of the GWAS for prognosis presented in Chapter 3; again, only the genotypic model was analysed by Cox regression. As there are no established predictive markers for the efficacy of 5-FU, and no data on MSI status available, the survival analysis was not adjusted for further variables.

Table 5-6 Predictive SNPs in PETACC 3

SNP	Chr	Position	HR (95% CI)	p-value
rs4715476	6	54730243	1.02 (0.85-1.23)	8.28e-01
rs9514816	13	107581122	0.93 (0.75-1.16)	5.08e-01

HR and 95% CI for the two SNPs that had been typed in the PETACC cohort.

Forty-seven SNPs fell below the threshold of p<1e-04. Of these SNPs, two had minor allele frequencies of less than 5% (rs8076116, rs11155360), and one SNP (rs273674) was not in HWE with p_{HWE}=1.11e-05 for the overall cohort. The remaining SNPs are presented in Table 5-5; five loci were tagged by two or more SNPs, and the redundant SNPs are not presented, leaving 38 SNPs. Two SNPs had been selected in the screen for prognostic markers presented in Chapter 3: rs4715476 on chromosome 6, 370kb from the tubulointerstitial nephritis antigen, and rs9514816 on chromosome 13, 77kb from DNA ligase 4. None of the SNPs were in or near genes known to be involved in the metabolism of 5-FU.

5.2.2 Further verification of predictive SNPs

All patients in PETACC 3 received 5-FU based chemotherapy, but data on survival were only available for the 30 SNPs included in the verification of prognostic SNPs, presented in section 4.1.1. Therefore only two of the 44 SNPs from section 5.2.1 above could be verified. The HR and 95% CI for the PETACC cohort are given in Table 5-6.

Table 5-7 Meta-analysis of the VICTOR and PETACC cohorts

SNP	HR (95% CI)	p-value	I^2	p_Q
rs4715476	1.24 (1.06-1.45)	7.15e-03	92.3	3.17e-04
rs9514816	1.18 (0.98-1.41)	7.51e-02	92.6	2.46e-04

A fixed-effects analysis was used for both SNPs.

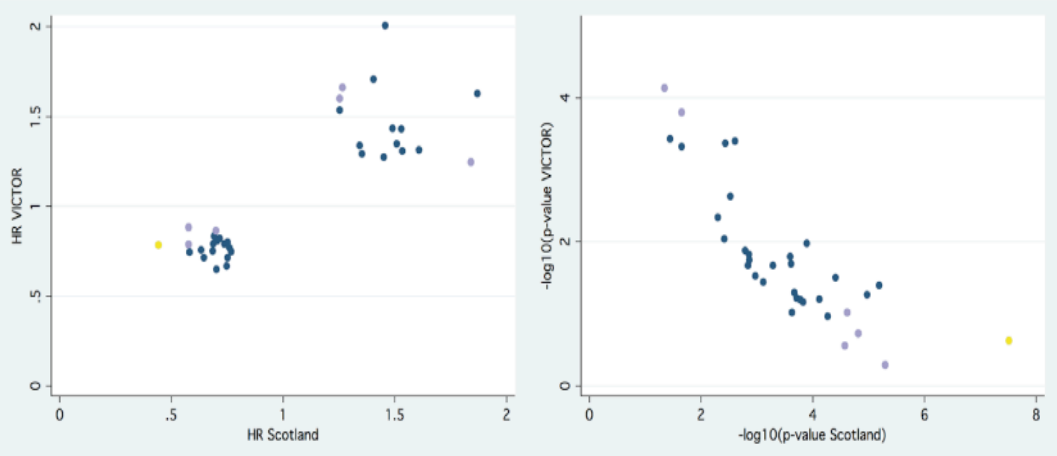
The estimates for these two SNPs were meta-analysed with those of the VICTOR cohort using a fixed-effects model (Table 5-7). Neither approached genome-wide significance, and the heterogeneity was highly significant and above 90% in both cases.

5.3 Discussion

The data presented in this chapter represent one logical next step following the analyses performed in Chapters 3 and 4: the low replication rate and the high proportion of recessive model SNPs that were included on the basis of an imprecise significance estimate provided by the logrank test. In addition, by adjusting for other prognostic variables in the initial screen, the risk of rendering a significant association insignificant once these are included in the model is avoided.

Perhaps the biggest advantage of the approach presented in this chapter is that by increasing probands, event rate and power, SNPs which in individual studies may only be borderline significant and would not have been taken into verification, are captured as well provided they are borderline significant in the other cohort as well. In fact, it is generally SNPs that have a similar HR and a significance level that is no more than two orders of magnitude apart that have met the overall threshold of p<1e-04, although the p-value for the most significant SNP, rs4074853, was driven predominantly by significance in the Scottish cohort, and did therefore not appear significant in a random effects model (Figure 5-1).

Figure 5-1 Plot of HR and -log$_{10}$(p-value) for VICTOR and Scottish cohort



For most SNPs, the HR correlate well between VICTOR and the Scottish cohort (r^2=0.72), including the most significant SNP, rs4074853 (yellow), although for HR<1, the Scottish cohort has lower HR while for HR>1, VICTOR has higher HR (left). The p-value distribution presented as the -log$_{10}$(p-value) is more spread along a crescent, with the median sum of the -log$_{10}$ from both cohorts being 5.1, showing that for most SNPs, the final significance level is driven by both SNPs, except for the ends of the crescent, where one SNP dominates, and the heterogeneity increases (lavender) and rs4074853 lies in that region (yellow).

This approach therefore largely selects SNPs after meta-analysis that may not have met the thresholds in the individual cohorts. Still, some of the top SNPs following meta-analysis are interesting, for example, rs1004361, lying within Kinase suppressor of ras 2 (*KSR2*), a scaffold protein required for KRAS-induced activation of the RAF-MEK-ERK phosphorylation cascade[322]. While the precise effect of *KSR2* is not yet known, *KSR1*, a close homolog, is required for cell cycle re-initiation following various types of DNA

damage[323]. rs1004361 would not have been taken forward from either cohort as the p-value in both was in the order of p=3e-03. A similar story emerges for rs2640882, within dermatan sulfate epimerase precursor, also known as *SART2*[324], a tumour-rejection antigen recognised by HLA-A24-restricted and tumour-specific cytotoxic T lymphocytes[325], where the individual p-values from VICTOR and Scottish cohorts would also not have met a p<1e-04 threshold. While the verification should not be based on selection by biological rationale, again based on the experience from Chapters 3 and 4, the existence of a strong biological rationale also does not make the SNPs less likely to replicate.

Contrasting these positives is the problem of different outcome measures, forced by available data. DFS was not available in the Scottish cohort and the event rate was low for CRC specific survival in VICTOR. As outlined in the introduction to this chapter, DFS and CRC specific survival are tightly correlated in 18 clinical trials looking at the effect of different treatment strategies[315]. Outcome in these trials will have been affected by germline pharmacogenomic variation and the type of therapy given as much as by the type of prognostic SNP screened for in this thesis. These studies are not an exact model for the analyses performed here, but nonetheless suggest that, in order to drive a more precise estimate of the risk contribution of individual SNPs, the comparison of two different CRC specific outcome measures may be valid.

The knowledge of prognosis is useful to plan the appropriate therapy, adjust existing treatment strategies, and help define subgroups for which new treatments are required. Knowledge of the likelihood of response to a given treatment, predictive information, will help to avoid giving ineffective systemic therapy, and therefore help improve the therapeutic ratio of currently available chemotherapy. An attempt was therefore made to perform a discovery phase for predictive marker using the patients of the VICTOR cohort who had received adjuvant chemotherapy, although no good verification cohort had been identified. No significant findings were made, and no SNP in or near genes of the metabolic pathway of 5-FU was in the top most significant SNPs. While the evidence for germline variation influencing outcome from 5-FU therapy is weak (section 1.6.2.1), it nonetheless exists. Consequently, either this evidence does not reflect the actual impact on 5-FU efficacy and was therefore not replicated, or, equally likely, the power to detect predictive germline markers was too small in the discovery phase presented here.

Compared to the screening cohorts used for CRC susceptibility loci, the cohorts used here remain small, between 30% and 40% of the susceptibility studies[44,299]. The relative event rates are smaller still, 19.43% in VICTOR and 32.15% in the Scottish cohort, compared to 49.20% and 48.48% CRC cases ('events') in the susceptibility screens, and an even

smaller cohort and event rate in the screen for predictive loci. Thus, despite the proven potential of GWAS to detect susceptibility loci for many diseases[39-41,44,326], the power to detect germline prognostic or predictive SNPs or loci is more limited. It is likely that GWAS for predictive markers will represent the next challenge beyond prognostic markers, as the cohorts needed for this approach require information not only on disease status and time to relapse, but also treatments given, ruling out most population based cohorts.

Ultimately, the success or otherwise of the approach taken in this chapter will depend on the ability to identify verification cohorts of sufficient size, with an appropriate ethnic and clinical background. Like for susceptibility GWAS, verification is key, and future efforts to define suitable cohorts for the verification of the SNPs identified in this chapter have already begun, and will ultimately be crucial for our understanding of the contribution of an individual's genetic make-up to prognosis and response to (systemic) therapy, a key prerequisite in the supposed age of personalised medicine.

Chapter 6

rs6983267 is associated with micrometastasis

Genome wide association studies have identified common, low-penetrance susceptibility alleles in a variety of cancers[38,39], and in CRC, 10 loci have now been described[42-46,52].

The high penetrance syndromes (FAP, HNPCC, JPS, PJS) are associated with an early age of onset, and early age of onset in turn has anecdotally been associated with worse prognosis in CRC[327,328]. Others have not found such an association[329,330], but not all studies performed a stage for stage comparison, and none split the patients into more than two age bands. While not well established in CRC, this is contrary to breast cancer, where young age of onset is an independent prognostic variable, possibly reflecting a different tumour biology[331]. It is therefore conceivable that the recently described low-penetrance CRC risk loci could predispose to an earlier age of onset and be associated with prognosis (Table 6-1).

Table 6-1 Known CRC low-penetrance risk loci

SNP	Alleles	Chromosome	OR (95% CI)	p-value	Gene	Distance
rs6983267[44]	T/G	8q24	1.21 (1.15-1.27)	1.3E-14	*MYC*	337Kb
rs16892766[42]	A/C	8q24	1.25 (1.19-1.32)	3.3E-18	*EIF3H*	27Kb
rs10795668[42]	A/G	10p14	0.89 (0.86–0.91)	2.5E-13	none	within 500Kb
rs3802842[52]	A/C	11q23	1.17 (1.12-1.22)	1.08E-12	*POU2AF1*	51Kb
rs4444235[46]	T/C	14q22	1.11 (1.08–1.15)	8.1E-10	*BMP4*	9.4Kb
rs4779584[43]	T/C	15q13	1.26 (1.19-1.34)	4.4E-14	*GREM1*	15Kb
rs9929218[46]	A/G	16q22	0.90 (0.87–0.94)	1.2e-08	*CDH1*	intronic
rs4939827[45]	T/C	18q21	0.85 (0.81–0.89)	1.0E-12	*SMAD7*	intronic
rs10411210[46]	T/C	19q13	0.83 (0.78–0.88)	4.6e-09	*RHPN2*	intronic
rs961253[46]	A/C	20p12	1.12 (1.08–1.16)	2.0E-10	*BMP2*	342Kb

OR are for CRC risk. *MYC* – *MYC* proto oncogene; *EIF3H* - eukaryotic translation initiation factor 3; *POU2AF1* - POU class 2 associating factor 1; *BMP* - bone morphogenic protein; *Grem1* - gremlin-1; *CDH1* - E-cadherin; *RHPN2* - Rho GTPase binding protein 2. Where more than one SNP has been described per locus, only the most significant SNP is listed.

6.1 Screening risk SNPs for association with outcome in VICTOR

All 10 SNPs were typed in the VICTOR cohort as part of the verification phase of the initial susceptibility GWAS. Four SNPs were typed by KASPar allele specific PCR (rs10758668, rs16892766, rs961253, rs10411210) with otherwise standard conditions for KASPar PCR (for primer sequences and conditions see Table 10-10 and Table 10-11). Genotypes for rs10411210 were generated by Kimberley Howarth. The remaining six SNPs were typed on the Illumina Hap300 arrays, and genotypes were generated for 947 patients (Table 6-2).

Not all samples had DNA available for KASPar genotyping, although overall, the call frequencies were high. Fidelity of the KASPar genotypes was assessed by direct sequencing within the laboratory. All SNPs were in HWE for the overall group as well as the relapsed and non-relapsed groups at the p>0.01 level.

Table 6-2 Genotypes in VICTOR

			Relapse			Non-relapse			
SNP	Total	Call rate	AA	AB	BB	AA	AB	BB	Method
rs6983267	945	99.8%	30	96	54	164	382	219	Illumina
rs16892766	809	100%	137	27	1	547	91	6	KASPar
rs10795668	892	99.0%	20	76	77	53	300	366	KASPar
rs3802842	947	100%	92	71	17	368	313	86	Illumina
rs4444235	946	99.9%	54	80	46	208	389	169	Illumina
rs4779584	947	100%	7	66	107	45	230	492	Illumina
rs9929218	947	100%	17	71	92	55	308	404	Illumina
rs4939827	946	99.9%	57	90	32	233	384	150	Illumina
rs10411210	754	95.2%	1	26	118	5	86	518	KASPar
rs961253	800	99.3%	18	73	56	83	320	250	KASPar

To test the influence on survival, all SNPs were analysed for the genotypic, recessive and dominant models by Cox regression in the VICTOR cohort in a univariate model. The cut-off for survival data for this analysis was 29 September 2008. Of the 10 SNPs, only rs6983267 and rs10768558 reached significance at the p<0.05 level (Table 6-3).

For rs6983267 in the recessive model, HR=0.66 (95% CI 0.44-0.99, p=4.57e-02), and for rs10768558 in the genotypic and recessive models, HR_{gen}=1.30 (95% CI 1.03-1.64, p=2.43e-02) and HR_{rec}=1.65 (95% CI 1.04-2.64, p=3.49e-02).

The significant associations were tested in multivariate analysis with stage. rs10768558 retained significance at the p<0.05 level for both models, while rs6983267 did not quite reach significance in the recessive model (p=9.10e-02, Table 6-4).

Table 6-3 HR for death for CRC risk SNPs

SNP	Model	HR (95% CI)	p-value
rs6983267	genotypic	0.84 (0.68-1.04)	1.18e-01
rs16892766	genotypic	1.10 (0.76-1.60)	6.18e-01
rs10795668	genotypic	1.30 (1.03-1.64)	2.43e-02
rs3802842	genotypic	0.88 (0.70-1.10)	2.58e-01
rs4444235	genotypic	0.99 (0.80-1.22)	9.17e-01
rs4779584	genotypic	1.07 (0.84-1.36)	6.01e-01
rs9929218	genotypic	1.09 (0.87-1.38)	4.53e-01
rs4939827	genotypic	0.91 (0.73-1.12)	3.75e-01
rs10411210	genotypic	1.16 (0.79-1.71)	4.56e-01
rs961253	genotypic	1.01 (0.79-1.29)	9.28e-01
rs10795668	recessive	1.65 (1.04-2.64)	3.49e-02
rs6983267	recessive	0.66 (0.44-0.99)	4.57e-02

HR and 95% confidence intervals determined by Cox regression in univariate analysis. Significant associations at the p<0.05 level are shaded in grey; for those SNPs not showing an association at this level in any model, only the genotypic model is given.

To test if the association of rs6983267 with DFS in univariate analysis was because of an association with stage, the genotypic and dominant models were analysed by logistic regression. For the recessive model, this was significant, OR_{rec}=0.68 (95% CI 0.49-0.93, p=1.68e-02), and for the genotypic model of borderline significance (Table 6-5). None of the models tested revealed a significant association of rs10768558 with stage (data shown for genotypic model only in Table 6-5.

Table 6-4 Multivariate Cox regression including stage

SNP	Model	HR (95% CI)	p-value
rs10795668	genotypic	1.27 (1.01-1.60)	3.84e-02
rs10795668	recessive	1.61 (1.01-2.57)	4.55e-02
rs6983267	recessive	0.70 (0.47-1.06)	9.10e-02

While the association of rs6983267 with stage is interesting, it is also only of borderline significance, and in itself does not offer any new prognostic information. If it predisposed to micro-metastatic disease (the definition of stage 3 CRC if metastasis is confined to lymphnodes on pathological examination), it would offer new prognostic information: pathological stage 2 patients who had undetected micro-metastasis should be offered adjuvant chemotherapy analogous to stage 3 patients.

Table 6-5 Association with stage

SNP	Model	OR (95% CI)	p-value
rs6983267	genotypic	0.83 (0.69-1.00)	5.11e-02
rs6983267	recessive	0.68 (0.49-0.93)	1.68e-02
rs10795668	genotypic	1.09 (0.89-1.34)	3.99e-01

There was a significant association of rs6983267 with stage in the recessive model, this was borderline in the genotypic model; for rs10795668 no association was found, data only shown for genotypic model.

Therefore, a new group with (micro)-metastatic disease was defined: those with stage 3 and hence confirmed metastatic disease on pathological examination, and relapsed stage 2 disease, who, on the assumption of adequate surgery and negative resection margins, must have had micro-metastatic disease in order to relapse. This group was compared to stage 2 patients who did not relapse, and are patients who did not have micro-metastatic disease.

The association of rs6983267 with the new category of (micro)-metastatic disease was significant in the genotypic and, in particular, in the recessive model, HR_{rec}=0.65 (95% CI 0.47-0.89, p=7.58e-03). This association remained if chemotherapy was included in the model (HR=0.59, 95% CI 0.40-0.86, p=6.93e-03). There was no association between this clinical group and rs10768558 in any model (Table 6-6).

Table 6-6 Association with micro-metastatic disease

SNP	Model	OR (95% CI)	p-value
rs6983267	genotypic	0.81 (0.67-0.97)	2.38e-02
rs6983267	recessive	0.65(0.47-0.89)	7.58e-03
rs10795668	genotypic	1.20 (0.97-1.48)	8.73e-02

There was a significant association of rs6983267 with micro-metastatic disease in the genotypic and recessive models, for rs10795668 no association was found in any model, data only shown for genotypic model.

6.2 Verification in further cohorts

The intriguing hypothesis raised by this work warranted further investigation, especially due to the retrospective nature of the initial analyses. Therefore, rs6983267 and rs10768558 were typed in further cohorts to test the association of rs10758668 with prognosis, test the association of rs6983267 with metastatic disease and stage, and refute its association with prognosis.

Table 6-7 Further genotyping for rs6983267 and rs10758668

Cohort	rs6983267	rs10758668
Australia	Collaborators	not typed
Quasar 1	Collaborators	not typed
Quasar 2	Illumina	not typed
Denmark	KASPar	KASPar
Epicolon	Collaborators	not typed
Finland	Collaborators	Collaborators
PETACC	Collaborators	Collaborators
VICTOR	Collaborators	No further samples

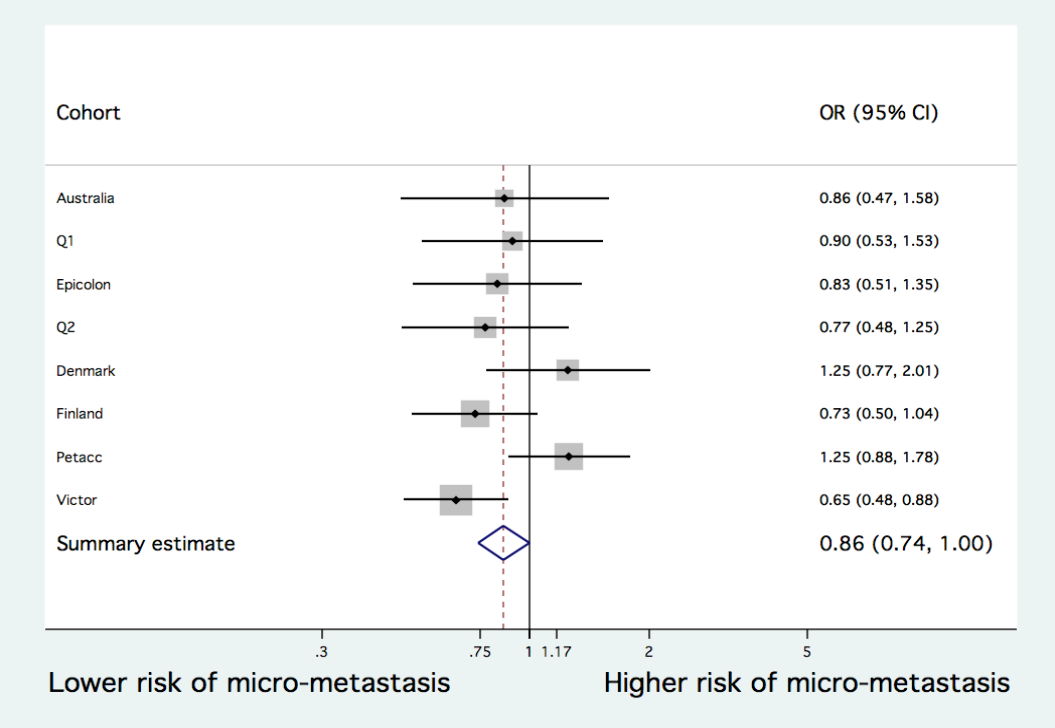
Genotyping work undertaken to further evaluate rs6983267 and rs10758668 in additional cohorts. Conditions for KASPar genotyping were as described above, all Quasar 2 samples were typed on Hap370 arrays. For the VICTOR cohort, there were an additional 185 samples with genotypes for rs6983267 available.

The required genotyping for rs10758668 and rs6983267 was undertaken by my collaborators, except for the Danish cohort, where both were typed by KASPar PCR (for primer sequences and conditions see Table 10-10 and Table 10-11), and rs6983267 in Quasar 2, typed on Illumina Hap370 arrays. This is summarised in Table 6-7.

6.2.1 rs6983267 and micro-metastatic disease

Genotype data for rs6983267 was available in a further seven cohorts: Australia, Denmark, Epicolon, Finland, PETACC, Quasar 1 and Quasar 2, providing genotypes for 4991 patients of all stages, including an additional 185 patients from the VICTOR trial who were genotyped from FFPE tissue by Dr Oliver Sieber. Follow-up was updated for the VICTOR cohort , the cut-off date was 29 January 2009. All cohorts were in HWE for the overall, non-relapsed and relapsed groups at the p>0.01 level.

Figure 6-1 Forest plot of association with micro-metastasis

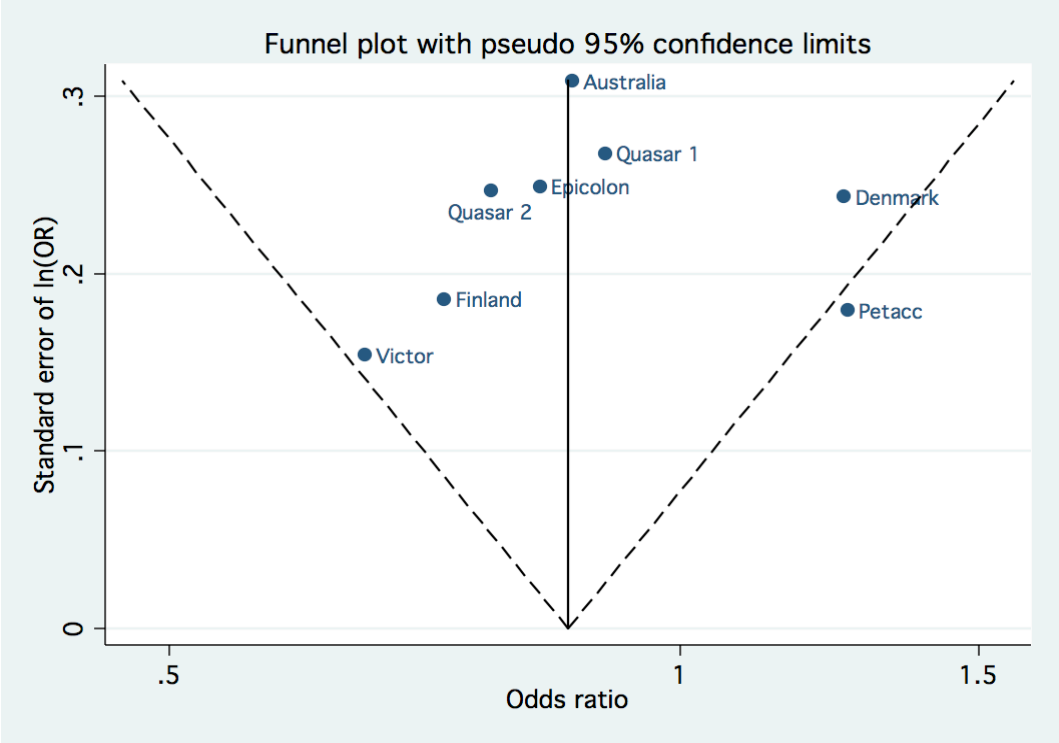


In the recessive model, TT is compared with TG/GG combined; all but two studies are suggestive of a lower risk of micro-metastasis associated with the T allele. Studies are ordered by increasing inverse variance weight.

The association of rs6983267with micro-metastatic disease in loco-regional CRC was tested in the eight cohorts available, and except in two cohorts (Denmark and PETACC), the direction of the effect was consistent with a decreased risk of micro-metastatic disease for minor allele homozygotes, although no other cohort reached statistical significance. The result from the VICTOR trial remained significant with longer follow-up. The combined OR of the risk of having micro-metastasis at surgery was statistically significant with OR_{rec}=0.86 (95% CI 0.74-1.00, p=4.33e-02) in the recessive model (Figure 6-1) with

no evidence of significant inter-study heterogeneity (I^2=36.8%, p_Q=0.135). In the genotypic model, which had had the smaller effect in the discovery phase in the VICTOR samples, the association was also significant OR_{gen}=0.92 (95% CI 0.85-1.00, p=4.10e-02).

Figure 6-2 Funnel plot of association with micro-metastasis



Funnel plot of the recessive model with pseudo 95% CI represented by the dotted lines. There is no evidence of significant bias in the studies analysed.

Visual inspection of the funnel plot (Figure 6-2) did not reveal any significant heterogeneity, and Egger's test was non-significant (p=0.333). To assess the impact of individual studies on the overall effect size, a leave one out analysis was performed: no study dominated the summary estimate.

The same analysis was performed, restricted to stage 1 and 2 patients who would be candidates for adjuvant chemotherapy if they had micro-metastasis (the equivalent to stage 3 if micro-metastases were detected on pathological examination of the lymphnodes). In the absence of adjuvant chemotherapy, this analysis is equivalent to predicting relapse in stage 1 and 2 patients. While the summary estimate did not reach formal statistical significance (OR_{rece}=0.86, 95% CI 0.69-1.08, p=1.93e-01; I^2=0%, p_Q=0.898) the result was consistent with an effect of the same magnitude and direction as before (Table 6-8).

Table 6-8 Association of rs6983267 with micro-metastatic disease in stage 1 and 2

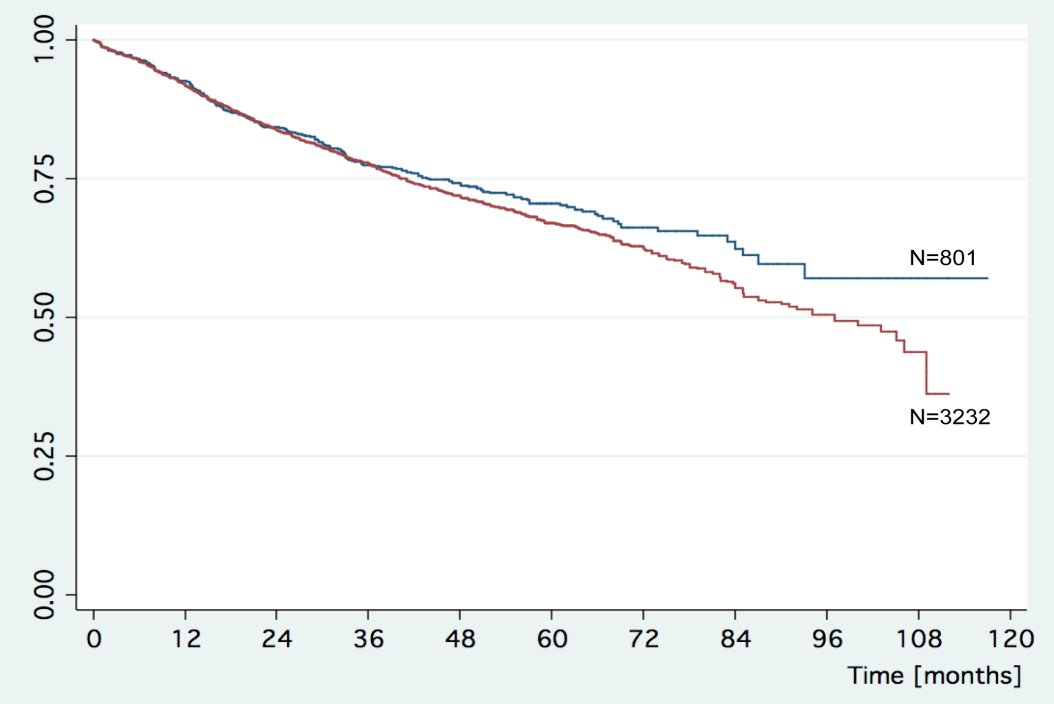
Cohort	OR (95% CI)	p_{rec}	p_Q
Australia	0.90 (0.23-3.45)	8.75e-01	
Quasar 1	0.86 (0.50-1.47)	5.80e-01	
Quasar 2	1.27 (0.24-6.67)	7.87e-01	
Denmark	1.15 (0.64-2.08)	6.33e-01	
Epicolon	0.99 (0.55-1.75)	9.65e-01	
Finland	0.73 (0.49-1.10)	1.29e-01	
PETACC	1.01 (0.50-2.08)	9.78e-01	
VICTOR	0.66 (0.35-1.25)	1.87e-01	
Summary estimate	0.86 (0.69-1.08)	1.93e-01	0.898

Data shown for the recessive model, restricted to 2,918 stage 1 and 2 patients.

6.2.2 rs6983267, survival and stage

None of the seven additional cohorts showed a significant association with survival, but in five, the results were compatible with a better prognosis for the TT genotype.

Figure 6-3 Kaplan-Meier curve for all patients



Kaplan-Meier curve for rs6983267 for all patients combined, recessive model: TT - navy, TG/GG - maroon. There is no statistical difference between the curves p=0.111.

When these data were meta-analysed, the overall effect was not significant with HR=0.93, 95% CI % CI 0.87-1.01, p=7.67e-02. The association with stage was tested, and the same picture emerged: no study was individually significant, neither was the overall effect with OR=0.93, 95% CI % CI 0.85-1.01, p=8.10e-02. There was no association when either survival or stage was tested in the recessive model, initially the

strongest association (Table 6-9), although there was an impression from the Kaplan-Meier curves that the two groups could be different (Figure 6-3).

Table 6-9 rs6873267 is not associated with survival or stage

Outcome	Model	HR (95% CI)	p_{gen}	p_Q
Survival	Genotypic	0.93 (0.87-1.01)	7.67e-02	0.646
Stage	Genotypic	0.93 (0.85-1.01)	8.10e-02	0.132

HR for survival, OR for stage.

However, when the event rate (either relapse or death) was analysed by stage, a relationship between stage and TT homozygosity emerged: the lower the stage, the lower the OR, 0.79 for stage 1 and 2.78 for stage 4. Within each stage group, the effect of rs6983267 was not significant (Table 6-10).

Table 6-10 T allele is inversely associated with event rate in early stage CRC

Stage	OR (95% CI)	p_{rec}	p_Q
1	0.79 (0.43-1.40)	4.27e-01	0.194
2	0.84 (0.66-1.07)	1.56e-01	0.989
3	0.91 (0.74-1.13)	3.92e-01	0.295
4	2.78 (0.82-14.64)	1.63e-01	0.847

The protective effect of the TT homozygote decreases with increasing stage.

6.2.3 rs10795668 and prognosis

rs10795668 was typed by allele specific PCR in three further cohorts: Denmark , Finland, and PETACC. All three cohorts were in HWE for the overall, non-relapsed and relapsed group at the p>0.01 level. No model reached statistical significance in any of the additional cohorts, although the data were consistent with an increased risk associated with the A allele. Updated data were used for the VICTOR cohort (data cut-off 28 January 2009), which slightly diminished the effect size without losing significance.

When meta-analysed, the summary estimate was not significant in the genotypic model with HR_{gen}=1.08 (95% CI 0.99-1.17, p=8.80e-02) with only moderate heterogeneity (Table6-11). There was no significant association in any other model, nor when only stage 2 and 3 patients were considered.

Table 6-11 rs10795668 and outcome

Cohort	HR (95% CI)	p_{gen}	p_Q
Victor	1.28 (1.02-1.6)	3.30e-02	
Danish	1.19 (0.95-1.49)	1.24e-01	
Finnish	1.03 (0.9-1.18)	7.11e-01	
PETACC	1.01 (0.86-1.18)	8.99e-01	
Summary estimate	1.08 (0.99-1.17)	8.80e-02	0.246

Data shown for the genotypic trend model, GG vs. GT vs. TT.

6.3 Discussion

The data presented in this chapter suggest that the T allele of rs6983267 protects from early micro-metastatic disease in CRC, thus resulting in disease that would be cured by surgery alone, not warranting further systemic adjuvant therapy. In addition, the T allele was also the protective allele in the GWAS for CRC susceptibility, meaning that TT homozygotes have not only a lower risk of CRC[44], but also are more likely to have a less aggressive form of the disease. Conversely, carriers of one or two G alleles appear more likely to develop CRC and have a more aggressive cancer. This association with tumour biology is not found with any of the other CRC susceptibility SNPs described so far, although there is some evidence that rs3802842 confers a higher risk of distal CRC[52].

The data appear real, as genotyping was robust and yielded similar results in six out of eight studies despite differences in genotyping method. The two studies with OR suggestive of a protective effect of the G allele (Denmark and PETACC) utilised KASPar genotyping and Sequenom technology, respectively, and the allele frequencies were similar in all cohorts, with T being the minor allele in line with previously published figures[44].

The distinction between metastasis and no metastasis points to a fundamental difference in cancer cell biology. Non-malignant cells do not leave the environment of like cells, and the ability to metastasise is one of the hallmarks of cancer. This ability in the early stages could have a significant impact on prognosis, but is less likely to be a major prognostic factor in the later stages, especially stage 4, when the ability to metastasise is less important than overall cancer burden - which of course does include metastasis, but at this point other factors presumably govern the number and size of metastasis and prognosis.

Thus, the apparent disappearance of a protective effect against an event (relapse or death) of the TT genotype with increasing stage is supportive of this concept: the likelihood not to have metastasised in stage 1 and 2 CRC means that the pathological stage truly reflects the 'biological' stage, whereas in stage 3 disease most patients already receive chemotherapy and in stage 4 disease obvious and often widespread metastasis are present.

It was not possible to link the propensity to early metastasis, or lack thereof, to a clinically useful outcome measure, and this finding remains one of scientific interest without at present impacting the management of patients with CRC. A separate marker for micro-metastatic disease would be clinically most useful in situations where there is no evidence of metastasis, namely in stage 1 and 2 CRC: patients at high risk of micro-metastasis could be offered adjuvant chemotherapy, while those patients at low risk could be reassured and spared the toxicity of chemotherapy, perhaps even in the presence of

adverse features which at present might otherwise convince clinicians to offer adjuvant therapy. Unfortunately, the data did not allow a firm conclusion regarding the impact on relapse rates: both for the dichotomous (relapse yes/no) as well as the time-to-event analysis, the magnitude of the effect size and its direction was the same in stage 1 and 2 patients as it had been when all patients are included, but did not reach statistical significance. It is not clear why this should have been, but variability of follow-up and associated clinical information may have played a part. Follow-up for the Danish and Finnish cohorts was ascertained through the national cancer registries, which only record death but not cause or relapse date. Furthermore, for these patients, no systemic treatment details were available, although it is likely that for both cohorts, stage 2 patients would not have received adjuvant therapy as they were ascertained prior to the general acceptance of adjuvant chemotherapy in high-risk stage 2 patients. Other potential confounders were the relatively low event rate in the VICTOR cohort, and the fact that only four cohorts were trial based. Furthermore, data on other prognostic features was not available, in particular no information regarding the resection margin, which, if positive, would indicate residual tumour predisposing to relapse without the 'need' for early occult metastasis from the primary tumour.

Again, rs6983267 stands out from other susceptibility SNPs in any cancer: not only is it the only SNP for which evidence of an effect on tumour biology exists, it is also the only SNP that predisposes to more than one tumour type[47], residing in the 8q24 region that more generally is associated with an increased risk of developing cancer[49]. The causal variants remain to be elucidated; the transcription factor *POU5F1P1* located 15Kb from rs6983267 is of an as yet undetermined functional significance. However, the location of rs6983267 on 8q24, 337Kb centromeric to the proto-oncogene *MYC*, invites speculation that changes in *MYC* expression and function could be the effector of both susceptibility and early metastasis, possibly via long-range cis-acting enhancer elements that are >250Kb from their respective gene in 50% of cases[308], a case now likely for CRC[332].

While *MYC* has not been specifically linked to the metastatic process, deregulation of *MYC* is one of the most common single gene abnormalities detected in human cancer[333], leading to occurrence and perpetuation of genomic instability which in turn plays an important role in the evolution of cancer cells[334], usually leading to the formation of distant metastasis. In B-lymphocytes, *MYC* expression decreases expression of LFA-1, involved in homotypic cell adhesion[335], and *MYC* could therefore also play a role in the release of CRC cells from the primary site to form metastasis.

MYC has been implicated in *WNT* signalling, being a downstream target of activated β-catenin, which jointly with TCF4 enhances *MYC* expression via the TCF4 binding site in

the *MYC* promotor[336], and deletion of *MYC* can rescue the abnormal intestinal phenotype associated with *APC* deletions[50]. It appears that the 8q24 region harbours a TCF4 binding site that may affect *MYC* expression[337], and there may be several tissue specific long-range enhancers of *MYC*, defined by the five LD blocks in this region, including one associated with CRC[49], possibly targeting only some of the many cellular processes (e.g. cell cycle[338], proliferation[339], apoptosis[340], cell adhesion[341]) in which *MYC* is involved. While there is no evidence linking *MYC* to prognosis in CRC, there is evidence from haematological malignancies that 8q24.21 rearrangements correlate with a more aggressive phenotype, speculatively the equivalent of early metastatic potential in solid organ malignancy, and a worse prognosis[342-344].

Chapter 7

rs6687758 is a novel CRC risk locus

rs6687758 is an A/G polymorphism on chromosome 1 at 220,231,571bp (1q41), with a MAF (G allele)=0.200 in HapMap samples of European ancestry.

In the recent GWAS for CRC susceptibility, it had been borderline significant at the p<0.05 level in one of two screening data sets (Scottish Phase 1, OR=1.16, 95% CI 1.00-1.35, p=0.049) but not the other (Corgi 2, OR=1.11, 95% CI 0.94-1.30, p=0.229) in the allelic model. The SNP was taken forward into two further screening sets for all SNPs at p<0.05 from Phase 1, where it was significant at the 0.05 level in both (Scottish Phase 2, OR=1.14, 95% CI 1.02-1.27, p=0.023; London Phase 2, OR=1.13, 95% CI 1.03 -1.24, p=0.009). Taken together, these cohorts suggested that rs6687758 is associated with CRC (p=3.39e-05), without reaching formal significance at the p=1e-07 level (Table 7-1). These analyses were performed by collaborators.

Table 7-1 Allelic OR for rs6687758 in four screening cohorts

Cohort	HR (95% CI)	P_{all}	p_Q
Corgi	1.11 (0.94-1.30)	2.29e-01	
Phase 2	1.13 (1.03-1.24)	8.54e-03	
Scotland Phase 1	1.16 (1.00-1.35)	4.90e-02	
Scotland Phase 2	1.14 (1.02-1.27)	2.30e-02	
Summary estimate	1.14 (1.07-1.21)	3.39e-05	0.975

OR and significance levels for rs6687758 in the screening cohorts for the GWAS for CRC susceptibility.

7.1 rs6687758 in VQ58 and Finnish samples

To confirm the initial findings, genotypes for two further datasets were collected: VQ58, a combined dataset of the samples of the VICTOR and Quasar 2 trials as cases, and the samples of the 1958 Birth Cohort as controls. The Finnish sample collection contained both cases and controls.

7.1.1 VQ58

This sample set consisted of 947 cases in the VICTOR trial, 494 cases in the Quasar 2 trial, and 1420 controls from the 1958 Birth Cohort. rs6687758 is on the Illumina Infinium Hap300 and Hap 550 arrays; it was typed as part of the GWAS for outcome (see Chapter 3), and the genotypes for the 1958 Birth Cohort were typed on the Illumina Hap550 arrays by collaborators as part of a GWAS for risk. The SNP performed well, with good cluster separation (GenTrain score=0.806 in VICTOR) and a call frequency of 99.7% in the VICTOR and Quasar 2 samples and 98.9% in the 1958 Birth Cohort (Table 7-2). The cohort was in HWE for all subgroups at the $p<0.05$ level.

The combined dataset showed a significant association between rs6687758 and CRC risk, OR=1.27, 95% CI 1.12-1.45, p=2.95e-04.

Table 7-2 rs6687758 genotypes in VQ58

Cohort	Type	Total	AA	AG	GG
Total cases	Cases	1436	875	493	68
• *VICTOR*		1. 944	2. 581	3. 314	4. 49
• *Quasar 2*		5. 492	6. 294	7. 179	8. 19
1958 Birth Cohort	Controls	1420	962	404	54

Genotypes for rs6687758 generated as verification for the findings of the London and Edinburgh Phases 1 and 2.

7.1.2 Finnish samples

The Finnish sample set consisted of 988 cases and 864 controls. The genotypes were generated using KASPar genotyping (see Table 10-12 and Table 10-13 for primers and condition). Of the 1852 samples, 65 were excluded due to unreliable sensors in the scanner and 22 samples failed genotyping, giving a call frequency of 98.7%.

Table 7-3 rs6687758 genotypes in Finnish cohort

Cohort	Type	Total	AA	AG	GG
Finland	Cases	928	476	385	67
Finland	Controls	803	437	317	49

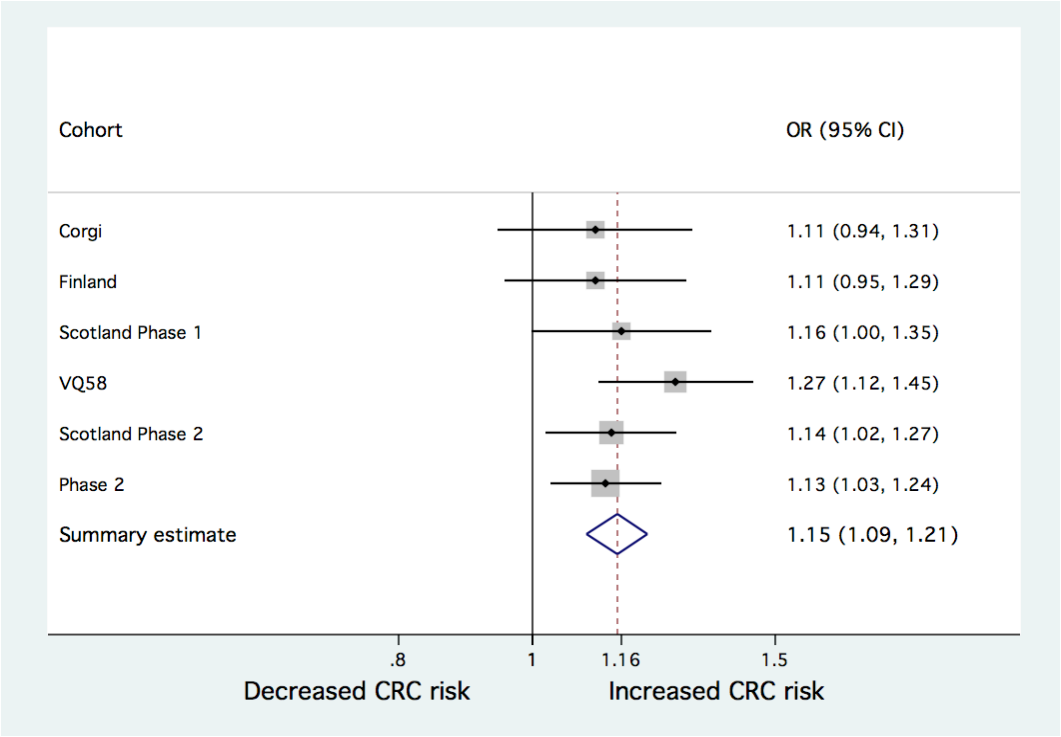
Reproducibility was tested with repeat analysis of 192 cases with very good concordance (98.96%). The discrepancy arose from one well on each plate that had failed initially being called in the repeat run. Genotypes were confirmed by sequencing a subset of samples, using the primers in Table 10-14 with standard conditions. The concordance was 100% in 83 samples for which genotypes were available from both techniques. Additional genotypes available after direct sequencing are included in the final analysis. The final genotype count is given in Table 7-3; the genotypes were in HWE at p=0.479.

There was no significant association between rs6687758 and CRC risk, OR=1.11, 95% CI 0.96-1.30, p=1.61e-01. Of note is that the OR, although not formally significant, is consistent with an effect of the same magnitude and direction as the other cohorts.

7.2 Meta-analysis of VQ58 and Finnish cohort

In order to derive a summary estimate of the CRC risk conferred by rs6687758, OR were generated for each group using logistic regression and pooled using a fixed effects model (Figure 7-1). The overall OR=1.15 (95% CI 1.10-1.21, p=5.04e-08) with no evidence of statistical heterogeneity (I^2=0%, p_Q=0.727). Because the frequency of the G allele was significantly different in the Finnish cohort ($MAF_{Finland}$=0.270) compared to all other cohorts (MAF_{mean}=0.203, range 0.199-0.209, p=1.5E-15), a random effects model was also used. The result of fixed and random models were identical, and are shown for the genotypic model (Table 7-4).

Figure 7-1 Forrest plot of association of rs6687758 with CRC risk



OR and 95% CI conferred by rs6687758. Studies are ordered by increasing inverse variance weight.

The overall effect size was also calculated using allelic, dominant and recessive models. All models retained significance for an association of the G allele with CRC risk, although the data are most consistent with a multiplicative model in which an increasing number of G alleles confers a constant increasing risk in a dose dependent manner: OR_{het1}=1.15 for the comparison of AA vs. AG and OR_{het2}=1.14 for the comparison of AG vs. GG. There was no evidence of a synergistic effect of two G alleles, OR_{hom} for the comparison of AA vs.

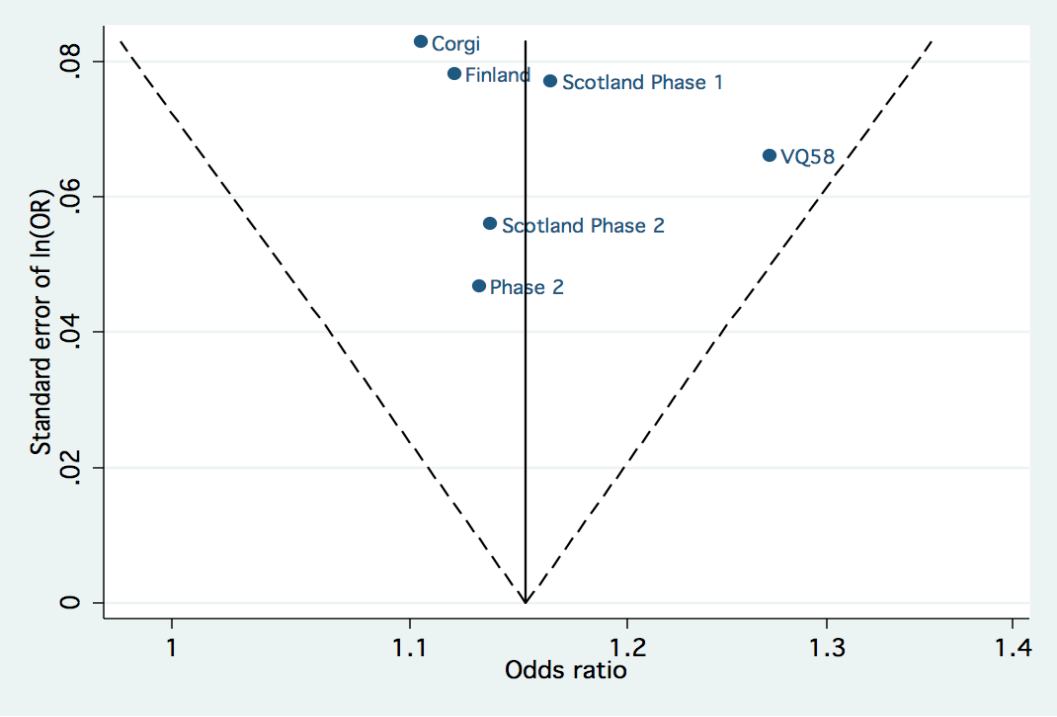
GG being again virtually identical to the product of OR_{het1} and OR_{het2} (OR_{het}=1.32), OR_{hom}=1.32 (95% CI 1.14-1.53, p=2.08e-04).

Table 7-4 Meta-analysis with different models

Model	OR (95% CI)	p-value	p_Q
Genotypic, fixed	1.15 (1.10-1.21)	5.04e-08	0.727
Genotypic, random	1.15 (1.10-1.21)	5.04e-08	0.727
Allelic	1.15 (1.09-1.21)	5.64e-08	0.678
Dominant	1.17 (1.10-1.24)	2.38e-07	0.521
Recessive	1.25 (1.08-1.45)	2.00e-03	0.386
Homozygous	1.32 (1.14-1.53)	2.08e-04	0.501

Statistical heterogeneity was further assessed using Egger's test (p=0.889) and visual inspection of a funnel plot (Figure 7-2). Neither revealed any evidence of statistical heterogeneity. To assess the impact of individual studies on the overall effect size, a leave one out analysis was performed. As expected the resulting OR was smallest if VQ58 , the study with the largest OR, was left out, but the result remained significant (OR=1.13, 95% CI 1.07-1.20, p=1.19e-05).

Figure 7-2 Funnel plot of association of rs6687758 with CRC risk



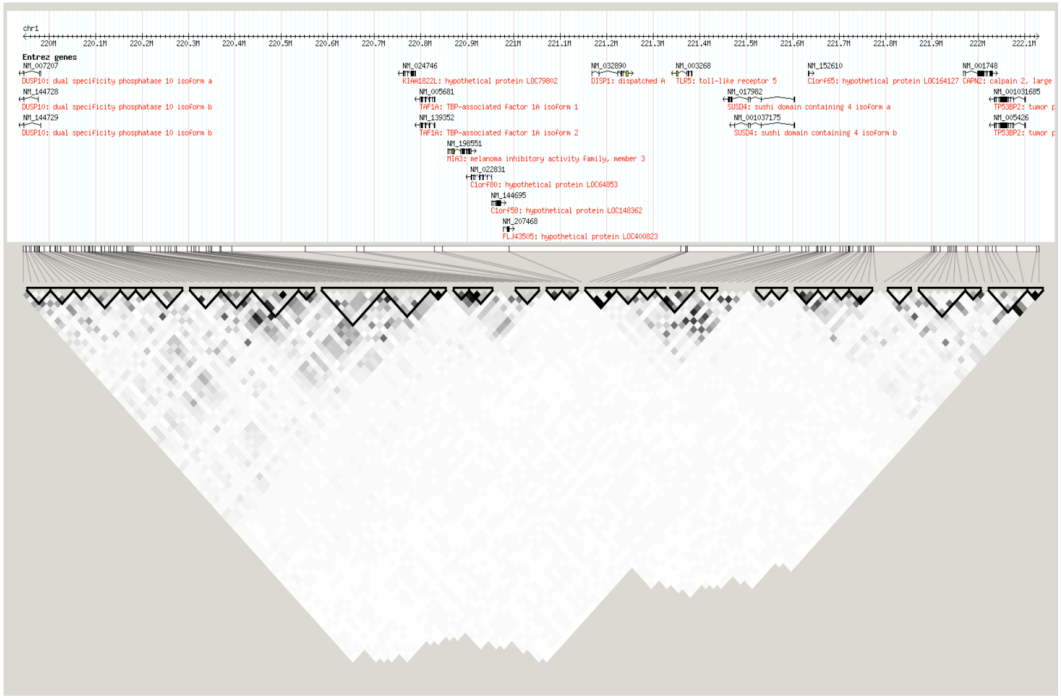
Funnel plot of the recessive model with pseudo 95% CI represented by the dotted lines. There is no evidence of significant bias in the studies analysed.

7.3 Functional considerations of rs6687758

There are no known genes in the immediate vicinity, rs6687758 lies 250kb upstream of and telomeric to dual-specificity phosphatase 10 (*DUSP10*), a known negative regulator of

the MAPK superfamily. Further telomeric lie TATA binding protein associated factor 1A (*TAF1A*, 566kb), melanoma inhibitory activity 3 (*MIA3*, 625kb), axin interactor (*AIDA*, 680kb), dispatched A (*DISP1*, 937kb), toll-like receptor 5 (*TLR5*, 1Mb), and tumour protein p53 binding protein 2 (*TP53BP2*, 1.7Mb).

Figure 7-3 LD blocks around rs6687758 defined by VICTOR genotypes

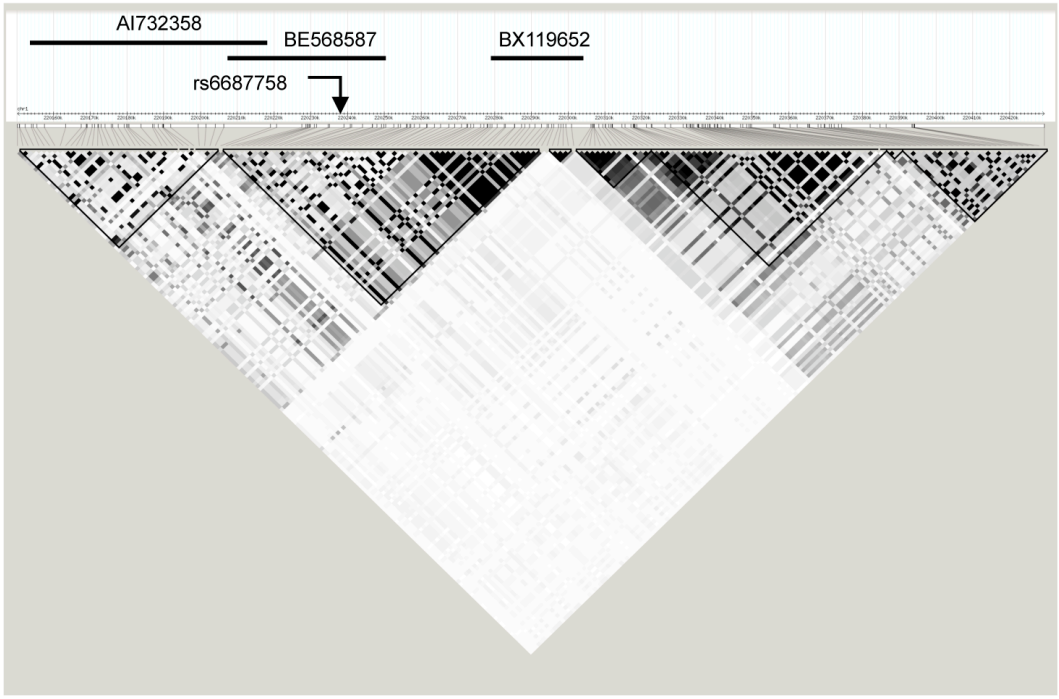


Long-range LD relationships between rs6687758 and *DUSP10* (chromosome 1, position 219.95Mb to 222.23Mb). There is evidence of a weak LD block extending towards *DUSP10*, there is no evidence of an LD block extending telomeric towards *TAF1A*.

To assess the potential for rs6687758 to have an impact on these genes, the region extending from DUSP10 to TP53BP2 was assessed using the genotypes for the VICTOR samples (Figure 7-3), likely to be more informative due to the higher number of probands than had been typed in HapMap. There is only a weak LD relationship within the region flanked by rs6687758 and *DUSP10*, but not extending telomeric towards *TAF1A*, *AIDA*, *TLR5*, and *TP53BP2*.

Further analysis of the immediate vicinity of rs6687758 revealed three ESTs within 60kb: *BE568587* from a carcinoma cell line, *AI732358* from colon tissue, and *BX119652* from MSI+ve colon cancer. The linkage relationship of this area was plotted for HapMap SNPs (chromosome 1, position 219.5Mb to 222.1Mb). Due to the higher number of SNPs typed, this gives a higher resolution (Figure 7-4). The LD block in which rs6687758 lies covers both *BE568587* and *BX119652*. There are no polyA tracts recorded in the UCSC genome browser for this region, although the NCBI UniGene site records a polyA signal for *BX119652*. The function of any these three ESTs remains unresolved.

Figure 7-4 LD blocks around rs6687758 defined by HapMap genotypes



LD blocks defined by HapMap genotypes (chromosome 1, position 220.16Mb to 220.43Mb). rs6687758 lies in an LD block with *BE568587* and *BX119652*, and to a lesser extend *AI732358*.

There are no other described transcripts, e.g. miRNA or ORFs, or functional elements in the region, e.g. long-range enhancers or CpG islands. There is minimal conservation in vertebrates across the region.

In an analogous analysis to that undertaken in Chapter 6 for published CRC risk loci, the impact of rs6687758 on DFS was assessed in the VICTOR data set. There was no evidence that rs6687758 influenced progression-free or overall survival in any of the genetic models (for DFS in genotypic model: HR=1.01, 95% CI 0.79-1.29, p=9.23e-01, see Figure 7-5), there was no association with stage (HR=0.98, 95% CI 0.79-1.22, p=8.79e-01) and micro-metastatic disease (HR=0.98, 95% CI 0.79-1.22, p=8.51e-01).

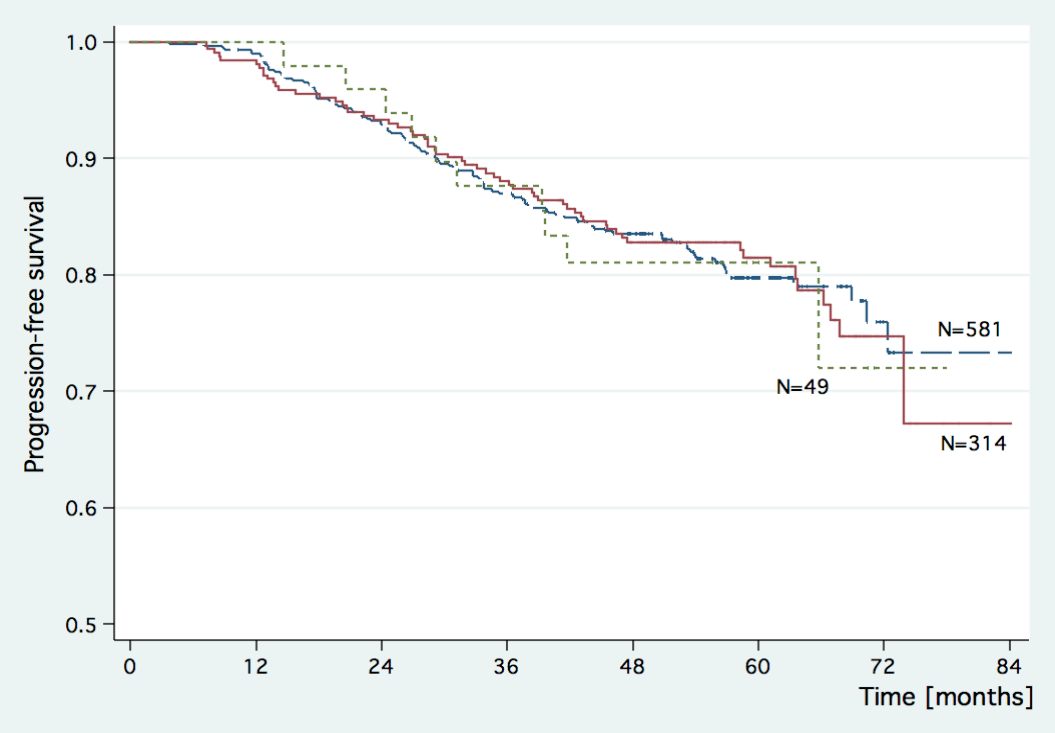
7.4 Discussion

The data presented here support rs6687758 as a further low-penetrance CRC susceptibility locus: the genotyping was robust, the effect size very similar in all cohorts with no statistical heterogeneity, and even when the study with the largest individual effect was omitted (VQ58), the resulting summary estimate remained very similar to the overall estimate (OR=1.13 vs. OR=1.15). The summary OR for all studies reached genome-wide significance with p=5.04e-08.

The MAF was virtually identical in the cohorts of Western European descent (MAF=0.203), but significantly different in the Finnish cohort (MAF=0.270), but the

resulting OR was the same in the Finnish cohort as it was in all but one of the European cohorts. Leaving the Finnish cohort out of the summary estimate did not change the overall OR and 95% CI appreciably. Rather than weaken the evidence, the different MAF in the Finnish samples is therefore suggestive of a real effect of rs6687758 on CRC risk, as even with a different allele distribution, the resulting OR is the same as in the Western European cohorts.

Figure 7-5 Kaplan-Meier curves for rs6687758



Kaplan-Meier curves for the three genotypes: AA – blue long dash, AG – red solid line, GG – green short dash. There is no difference between the survival curves (p=0.923).

rs6687758 had not been further evaluated in the initial meta-analysis of the four cohorts (Corgi, Phase 2, Scotland Phase 1 and Phase 2) of the GWAS for CRC susceptibility as the combined threshold did not reach the pre-specified level (p<1e-05)[46]. The power of that study to detect major common loci conferring quite large risks (1.2 or greater as was the case for rs6983267) was high and it is unlikely hat there are many additional CRC susceptibility SNPs with similar effects for alleles with frequencies >0.2 in populations of European ancestry. This underlines the efforts needed to detect low-penetrance variants that confer risks <1.2 and/or have MAF<0.2, as variants with these characteristics are likely to make up a large proportion of susceptibility loci for CRC, either because the effect sizes of the causal variant are truly that small, or the causal variant is being tagged inefficiently by the SNPs included in the arrays used for the initial screen. While the Illumina Hap550 arrays used in the London and Edinburgh Phase 1 are designed to capture the majority of common SNPs with MAF of >0.2 in European populations with

r^2>0.80, only around 10% of SNPs with MAF=0.05-0.1 are tagged at r^2>0.80, thus limiting the ability to detect this type of susceptibility SNP[46]. It is therefore plausible that there are many other loci with similar genetic characteristics to rs6687758 that remain to be discovered.

It is likely that other classes of genetic variation exist that can confer an increased CRC risk, a view supported by the continued excess of associations observed over those expected, but for which current strategies for GWAS are not ideally suited[46]. For susceptibility loci with lower MAF but higher penetrance, e.g. MAF=0.01 and a genotypic relative risk of 5 giving a relevant population attributable risk, tagging SNPs on currently available arrays are not well paced to detect an association (see also Chapter 8). In addition, the content of earlier generations of SNP array did not capture CNV well, and these may also affect CRC risk, but more recent arrays have started to address this issue by including dedicated CNV content, SNPs, and other markers, that are located in regions of known CNV. Whether the data generated from such loci is of sufficient quality to analyse CNV in a manner analogous to SNPs remains to be determined[345].

Expanding the scale of GWAS by including more SNPs in the initial screen, analysing CNV content, increasing sample size and the number of SNPs taken into the replication phase, with all the costs that this would involve, will likely identify additional susceptibility loci for CRC. In parallel with further expansion of GWAS, however, functional studies of the identified loci are important, as many do not offer an immediate explanation for their association with CRC risk[42], and even for rs6983267 located in a general cancer susceptibility region on chromosome 8q24, the functional correlate or causal variant has not been identified, despite the evidence pointing towards *MYC* (see section 6.3). Similar uncertainties remain for rs6687758, as no elements of any obvious functional relevance have been found in its vicinity, and while several ESTs are found relatively close by (60Kb), it is not clear whether they represent actual genes and what their function might be. *DUSP10* would make a good candidate but does not lie within an LD block with rs6687758, and would therefore require rs6687758 to tag an interference with long-range enhancers or chromatin folding as mechanisms for altering gene expression.

Chapter 8

Tagging SNPs tag potential cancer loci

There are two approaches to selecting content for the current SNP typing platforms: one is to select SNPs in a quasi random fashion, with an emphasis on regions of interest (e.g. Affymetrix[346]), or to choose SNPs based on their relationship to each other, i.e. tagging SNPs (e.g. Illumina). The latter approach improves the number of SNPs in LD at r^2>0.8 for any given SNP typed on the array from 69% on the Affymetrix Mapping 500K Array Set[346] to around 89% on the Illumina Hap500 arrays in Caucasian populations[347]. The LD relationships are based on the analysis of SNPs typed for the HapMap project, which aims to describe common variants (SNPs with MAF>0.05) throughout the genome, and estimates "*to capture untyped common variation with an average maximum r^2 of between 0.9 and 0.96 depending on population.*"[348].

Given that the SNP selection for Illumina arrays was performed using HapMap SNPs, genetic variation not captured by HapMap may be covered less well[348]. This includes non-HapMap SNPs, insertion/deletion changes and polymorphic repeats. To test whether such changes are tagged efficiently, loci with some prior rationale for being associated with CRC were typed in subsets of Corgi 2. The rationale for each is briefly described in the section on each marker.

The markers in this chapter on the whole are not SNPs; they therefore do not conform to the Illumina naming convention and all loci are represented as major and minor allele.

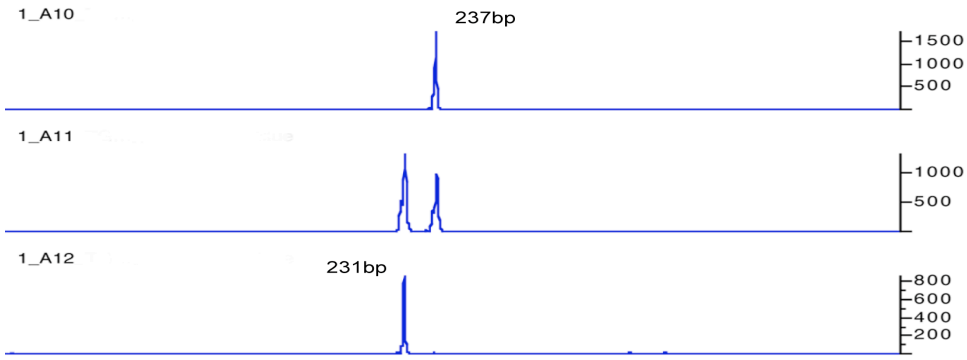
8.1 Screen for CRC risk in plausible candidates

This section details the outcome of the risk screen performed prior to assessing the LD relationship between these loci and SNPs present on the Illumina Hap500 arrays, typed in the Corgi 2 cohort as part of the GWAS for CRC susceptibility.

8.1.1 *TGFβR1*6A*

The TGF-β receptor 1 (*TGFβR1*) on chromosome 9 contains a polymorphic nine alanine stretch (*9A), and alleles containing 5, 6, 8, 10, 11, and 12 alanine residues (*5A, *6A, *8A, *10A, *11A, *12A) have also been described[349]. Of the variant alleles, *6A (rs11466445) is the most common, with up to 22.7% of probands carrying at least one allele[350], with a MAF between 5.0%-11.7% in published series[350,351]. *6A has previously been described as a cancer susceptibility allele for CRC, based on a meta-analysis of four datasets[352], although the statistical analysis was not prepared in accordance with generally accepted methods[353].

Figure 8-1 GeneScan for *TGFβR1* alleles



GeneScan trace for the *TGFβR1*9A and *6A alleles. Homozygote 9A/9A (top) with a product size of 237bp, heterozygote 9A/6A (middle), and homozygote 6A/6A with a product size of 231bp (bottom).

In an updated analysis of the original report, the same group, adding another two datasets to a total of 1585 CRC patients, again found an association with CRC[354] although these results have not been replicated by others[355].

In order to test the association of *TGFβR1*6A with CRC, the samples of the Corgi cohort were genotyped for *TGFβR1* alleles using GeneScan (Figure 8-1). The conditions for the PCR reaction and primers are given in Table 10-15.The genotypes were in HWE for overall, case and control cohorts, and are given in Table 8-1. There was no evidence for an association with CRC risk associated with carrying *6A (OR=1.04, 95% CI 0.83-1.31, p=7.06e-01).

Table 8-1 Genotypes for potential cancer variants

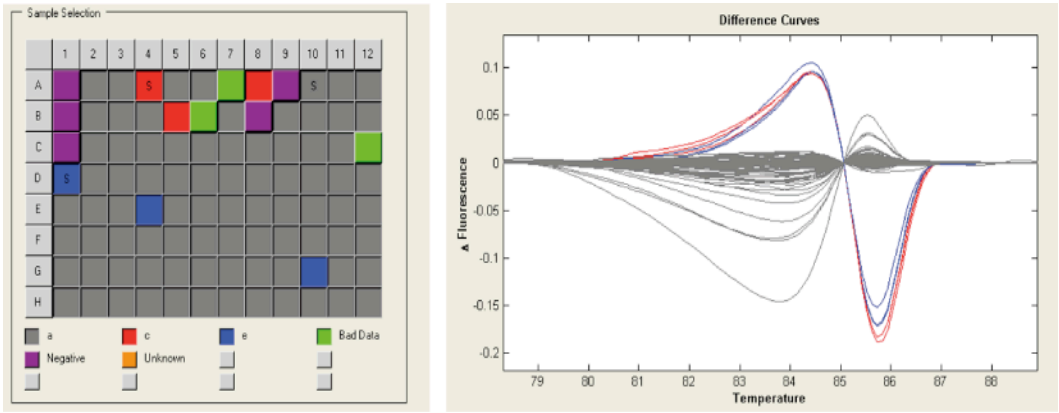
	Cases			Controls		
Marker	MM	Mm	mm	MM	Mm	mm
TGFβR1	673	145	10	746	159	8
APC E1317Q	816	14	0	905	16	0
rs28931588	155	0	0	162	0	0
rs28931589	155	0	0	162	0	0

M is the major allele, m the minor allele; for *TGFβR1* M=*9A, for *APC* E1317Q M=G.

8.1.2 APC E1317Q

Germline mutations in *APC* on chromosome 5 play a major role in inherited CRC. *APC* E1317Q (rs1801166) is a non-truncating mutation near the mutational hotspot at 1307, resulting in a substitution of the normal glutamate for a glutamine residue. This germline mutation has been associated with an increased risk in white Caucasian populations of CRC[356] and colorectal adenomata[357], a finding not replicated by others[358]. To test the possible association of *APC* E1317Q with CRC, the samples of the Corgi cohort were typed using the LightScanner methodology (Figure 8-2). The conditions for the PCR reaction and primers are given in Table 10-15.

Figure 8-2 LightScanner output for APC E1317Q alleles



Left panel shows distribution of samples in 96-well plate. Grey squares are *APC* wildtype; blue squares, positive controls (*APC* E1317Q carriers); red squares, carriers detected in the Corgi 2 cohort. Purple squares are blanks, and green squares, failed samples. The latter two categories are excluded from the curves shown in the right panel plotting changes in fluorescence in response to temperature changes during the LightScanner analysis. Colours are the same as in the left panel, and the difference between wildtype and carriers is evident.

The genotypes were in HWE for overall, case and control cohorts, and are given in Table 8-1. There was no evidence for an association with CRC risk associated with carrying the minor allele (OR=0.97, 95% CI 0.47-2.00, p=9.35e-01).

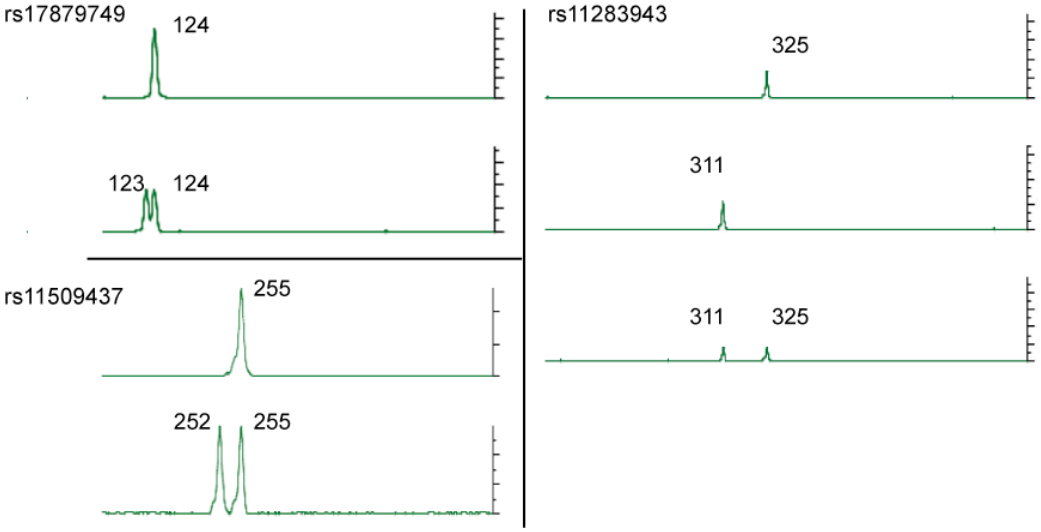
8.1.3 Insertion-deletion polymorphisms

The May 2004 assembly of the UCSC Genome browser was screened for non-synonymous insertion-deletion polymorphisms of known genes. Thirty-eight polymorphisms were selected based on known MAF >0.05, putative gene involvement in a cancer relevant function, and location of the polymorphisms within the middle 80% of the gene in an attempt to enrich for functionally relevant polymorphisms. All were screened in 192 cases and 192 controls of the Corgi 2 cohort using GeneScan (Figure 8-3). This design allowed for >80% power if the allele frequency was 5% in controls and 14% in cases, or 10% in controls and 21% in cases. Furthermore, if three or less alleles were

detected, there would be a >95% chance that even the complete Corgi set would not have sufficient power to detect an association and the polymorphism in question could be discarded from any future investigation for association with CRC risk.

The primers and conditions are given in Table 10-16.

Figure 8-3 GeneScan for insertion-deletion alleles



GeneScan trace for heterozygotes for rs17879749, rs11509437, and rs11283943 showing the difference of PCR product of 1bp, 3bp, and 14bp, respectively. The scale to the right of each lane is in arbitrary units determined by the analysis software.

Two polymorphisms failed genotyping: rs11478027 in TRPC4-associated protein and rs10667315 in the CD40 antigen. The remaining polymorphisms performed well on genotyping, with a median call rate of 97.5% (range 93.8%-99.7%). All but one polymorphism (rs5790928) were in HWE in the overall and control groups. In only 14 polymorphisms could the minor allele be detected, the remainder were monomorphic in both cases and controls. None of the 14 polymorphic variants was associated with an increased risk of CRC at the $p<0.05$ level, and only one polymorphism (rs11509437) located in glutathione-S-transferase omega 1 came close to statistical significance (OR=2.37, 95% CI 0.89-6.31, p=8.40e-02). The genotypes, associated genes and OR are given in Table 8-2.

Table 8-2 Genotypes, associated genes and statistical significance of insertion-deletion polymorphisms

					Controls			Cases				
SNP	Chr	Position	Gene	Alleles	AA	AB	BB	AA	AB	BB	OR (95% CI)	p-value
rs10697058	17	3753242	Purinergic receptor P2X1	-/TTTTTTTT	180	0	0	170	0	0		
rs11283943	5	112434687	Mutated in colorectal cancers	-/CGCGCTGT CTTCCT	83	73	21	85	69	19	1.08 (0.78-1.48)	6.57e-01
rs11350445	22	39883245	E1A binding protein p300	T/-	177	0	0	173	0	0		
rs11364237	2	112324604	Anaphase promoting complex subunit 1	T/-	177	0	0	172	0	0		
rs11376162	11	65129818	MAP kinase kinase kinase	-/G	179	0	0	165	0	0		
rs11424286	18	59477610	Serine proteinase inhibitor	-/T	175	1	0	173	0	0	2.96 (0.12-72.85)	5.07e-01
rs11476163	20	31728743	E2F transcription factor 1	G/-	169	0	0	173	0	0		
rs11477710	20	18419125	Retinoblastoma binding protein 9	T/-	163	0	0	172	0	0		
rs11509437	10	106012825	Glutathione-S-transferase omega 1	AGG/-	148	13	0	166	6	0	2.37 (0.89-6.31)	8.40e-02
rs11564598	19	53272844	Phospholipase A2, group IVC	C/-	169	10	0	159	5	0	1.86 (0.63-5.49)	2.63e-01
rs11564619	19	53262887	Phospholipase A2, group IVC	G/-	178	0	0	165	0	0		
rs140596	15	46563377	Fibrillin 1	A/-	175	2	0	174	0	0	4.94 (0.24-103.3)	3.03e-01
rs17001464	19	11549959	Acid phosphatase 5	G/-	162	0	0	172	0	0		
rs17879749	11	102172656	Matrix metalloproteinase 1	T/-	163	3	0	169	6	0	0.52 (0.13-2.11)	3.62e-01
rs2066730	1	183647677	Phospholipase A2, group IVA	T/-	177	0	0	172	0	0		
rs3069752	7	99601783	Paired Ig-like type 2 receptor beta	GGA/-	108	51	10	122	47	5	1.36 (0.92-2.00)	1.21e-01
rs3092768	20	39474243	Chromodomain helicase DNA binding protein 6	C/-	174	0	1	171	3	0	0.66 (0.11-3.98)	6.51e-01
rs3212987	19	50604348	CD3E antigen	TTC/-	93	68	14	86	78	9	0.98 (0.71-1.37)	9.25e-01
rs3214276	5	140753085	Protocadherin gamma subfamily A, 8	C/-	181	0	0	171	0	0		
rs3217458	19	1776929	Transcription elongation factor B	CTC/-	121	51	7	102	52	11	0.77 (0.53-1.11)	1.64e-01
rs4067742	2	112330488	Anaphase promoting complex subunit 1	-/TTA	175	1	0	175	0	0	2.99 (0.12-73.68)	5.03e-01

Table 8-2 Genotypes, associated genes and statistical significance of insertion-deletion polymorphisms (cont'd)

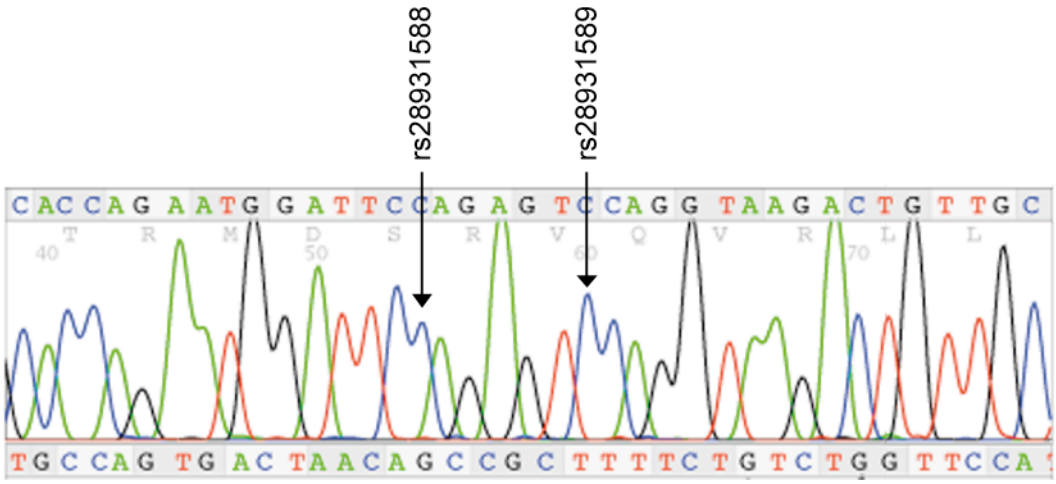
					Controls			Cases				
SNP	Chr	Position	Gene	Alleles	AA	AB	BB	AA	AB	BB	OR (95% CI)	p-value
rs4987226	11	59383178	Transcobalamin I precursor	A/-	178	0	0	174	0	0		
rs5030651	3	10158713	von Hippel-Lindau tumour suppressor	C/-	163	0	0	171	1	0	0.35 (0.01-8.64)	5.22e-01
rs5773188	1	27360299	Synaptotagmin-like 1	C/-	178	0	0	172	0	0		
rs5790928	11	33326090	Homeodomain interacting kinase 3	G/-	165	2	0	172	1	1	0.69 (0.12-4.17)	6.89e-01
rs5803440	13	47984938	Regulator of chromosome condensation	T/-	181	0	0	171	0	0		
rs5814559	15	91287210	Chromodomain helicase DNA binding protein 2	T/-	181	0	0	169	0	0		
rs5820291	17	34818392	PPAR binding protein	GTC/-	177	0	0	175	0	0		
rs5844947	22	30065477	zinc finger protein 278	C/-	175	0	0	173	0	0		
rs5848002	3	37489931	Integrin, alpha 9 precursor	C/-	177	0	0	166	0	0		
rs5848302	3	38019751	Villin-like protein	G/-	176	0	0	172	0	0		
rs5883925	7	44914210	Cell cycle progression 2 isoform 1	-/C	173	0	0	175	0	0		
rs5888463	7	151287527	Mixed-lineage leukaemia 3	C/-	176	0	0	175	0	0		
rs9332131	10	96699028	Cytochrome P450, family 2, subfamily C	A/-	168	0	0	174	0	0		
rs9332736	6	32010046	Complement component 2 precursor	GTGGACAGGGTCAG GAATCAGGAGTCTG/-	172	2	0	176	2	0	1.02 (0.14-7.30)	9.82e-01
rs9333357	9	21218001	Interferon alpha 17	T/-	181	0	0	171	0	0		

Genotypes and associated data for 36 insertion-deletion polymorphisms. Two polymorphisms failed genotyping (rs11478027 and rs10667315); these are not included in the table.

8.1.4 beta-catenin: rs28931588 and rs28931589

Several somatic mutations in the degradation targeting box of the β-catenin gene on chromosome 3 are associated with childhood hepatoblastoma (hepatocellular carcinoma, HCC)[359]. Two of these have been described as SNPs in the germline: Asp32Tyr (rs28931588) and Gly34Val (rs28931589). While to date there is no evidence in the literature associating these polymorphisms with CRC, they are plausible candidates due to the importance of *WNT* signalling in the development of CRC. These two variants were 192 cases and 192 controls. The power considerations were the same as those described in section 8.1.3. The genotypes were determined by direct sequencing (Figure 8-4) using the conditions given in Table 10-15, and the sequencing reaction performed with the reverse primer.

Figure 8-4 Sequencing trace for rs28931588/9 alleles



The arrows show the location of the two SNPs, rs28931588 and rs28931589, both are homozygous for the major allele G, read on the reverse strand as C.

There were no minor alleles detected, with all genotypes being major allele homozygotes (Table 8-1). rs28931588 and rs28931589 are not associated with CRC based on this samples set.

8.2 LD between variants screened and tagSNPs on Illumina Hap550

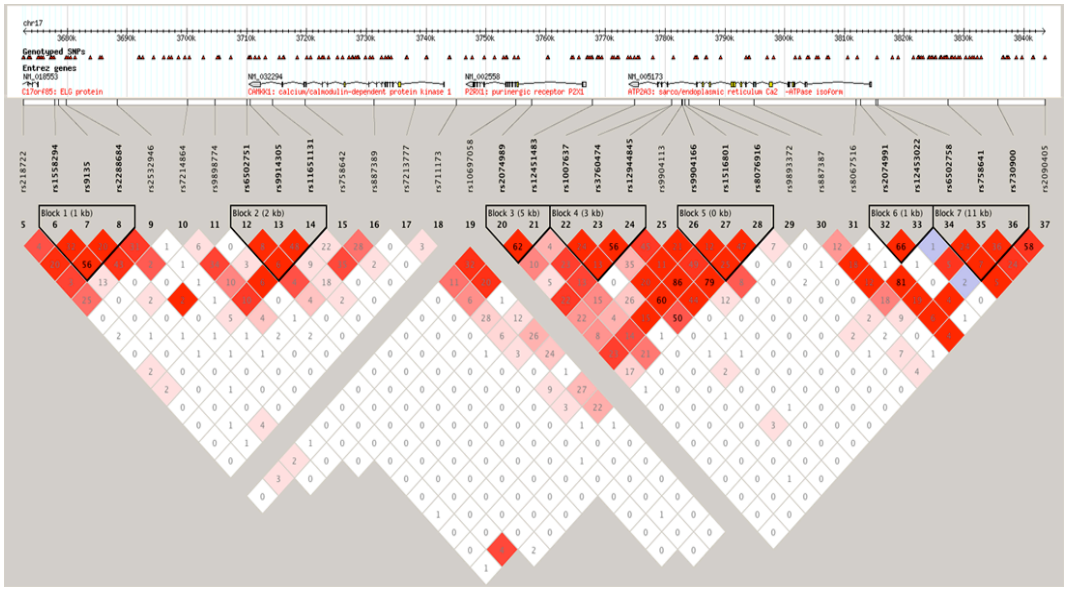
The samples of the Corgi 2 cohort had been typed on Illumina Hap550 arrays as part of the GWAS for CRC risk, and the LD relationships between the polymorphisms typed in section 8.1 and the tagSNPs of these arrays were evaluated. For each polymorphisms, tagSNPs 100Kb either side were extracted, and the maximal r^2 and *D'* were determined using Haploview 4.1.

As expected, it was only possible to generate r^2 or *D'* values for the non-monomorphic variants, monomorphic variants (rs28931588/9 and 22 insertion-deletion polymorphisms) did not have r^2 and *D'* values for any of the tagSNPs and thus they are not tagged by

SNPs on the Hap550 arrays (see Figure 8-3 for an example of the LD relationship in such a case). The remaining 16 polymorphic variants (*TGFβR1**6A, *APC* E1317Q, and 14 insertion-deletion polymorphisms) were tagged efficiently ($r^2 \geq 0.80$) only if the variant in question had MAF≥0.05. For variants with MAF<0.05, the maximum $r^2=0.12$, while *D'*=1 for all but two polymorphism, for *APC* E1317Q, *D'*=0.914 and for rs5790928, *D'*=0.625 (Table 8-3). For the five variants with MAF>0.05 (TGFβR1, rs11283943, rs3069752, rs3212987, rs3217458), r^2 ranged from 0.82 to 0.99, and *D'* from 0.980 to 1.

Of the insertion-deletion polymorphisms that are likely to be most deleterious to the resulting protein, i.e. those that do not change the base pair sequence by multiples of three, only one polymorphism was common (rs11283943, MAF=0.317), and only one further had a MAF>0.01 (rs17879749, MAF=0.013). The remaining common polymorphisms all inserted or deleted whole codons.

Figure 8-5 LD block surrounding rs10697058



rs10697058 (SNP 19) is monomorphic and not tagged by any of the SNPs on the Illumina Hap550 arrays. This plot is representative of other monomorphic SNPs.

For variants with MAF<0.05, in the cases where there was a high *D'*, this gave a misleading picture about the ability to tag the polymorphism as the minor allele was not tagged well: for *APC* E1317Q, the major allele of its tagSNP, rs10515449, is associated with the proposed CRC risk allele in 6.4% of cases, while the major allele is associated with the *APC* wildtype allele in 93.6%. For the minor allele of rs10515449, it is associated with proposed risk allele in 0.1% and the wildtype allele in 99.9%. Neither allele of rs10515449 therefore gives a good estimate of the chance of the proband carrying *APC* E1317Q when only the former is typed. Conversely, for *TGFβR1**6A with MAF=0.098,

the major allele of its tagSNP, rs7850895, captures the wildtype allele of *TGFβR1* in 98.5%, and *TGFβR1**6A is captured by the minor allele of rs7850895 in 98.8%.

Table 8-3 Tagging of insertion-deletion polymorphisms

SNP	Δ	M	m	MAF	r^2	*D'*	tagSNP	MAF	MM	Mm	mM	mm
rs10697058	8	700	0	0	0	0						
rs11283943	14	478	222	0.317	0.993	1	rs2227950	0.315	68.1	0.1	0	31.7
rs11350445	1	700	0	0	0	0						
rs11364237	1	698	0	0	0	0						
rs11376162	1	688	0	0	0	0						
rs11424286	1	697	1	0.001	0.014	1	rs3853683	0.090	91.0	8.9	0	0.1
rs11476163	1	684	0	0	0	0						
rs11477710	1	670	0	0	0	0						
rs11509437	3	647	19	0.029	0.054	1	rs156697	0.355	64.9	32.3	0	2.9
rs11564598	1	671	15	0.022	0.066	1	rs1019786	0.254	74.6	23.2	0	2.2
rs11564619	1	686	0	0	0	0						
rs140596	1	700	2	0.003	0.001	1	rs9806163	0.234	76.1	23.6	0.3	0
rs17001464	1	668	0	0	0	0						
rs17879749	1	673	9	0.013	0.012	1	rs1010698	0.476	51.5	47.2	1.3	0
rs2066730	1	698	0	0	0	0						
rs3069752	3	558	128	0.187	0.990	1	rs6955367	0.192	81.1	0.1	0	18.7
rs3092768	1	693	5	0.007	0.007	1	rs4142393	0.490	49.7	49.6	0.7	0.0
rs3212987	3	504	192	0.276	0.874	1	rs3212986	0.298	69.7	2.6	0.1	27.5
rs3214276	1	704	0	0	0	0						
rs3217458	3	549	139	0.202	0.982	1	rs3746037	0.202	79.5	0.3	0	20.2
rs4067742	3	701	1	0.001	0.003	1	rs9653414	0.325	67.1	32.8	0	0.1
rs4987226	1	704	0	0	0	0						
rs5030651	1	669	1	0.001	0.012	1	rs1642739	0.111	88.8	11.1	0	0.1
rs5773188	1	700	0	0	0	0						
rs5790928	1	677	5	0.007	0.008	0.63	rs3847595	0.271	72.6	26.7	0.2	0.5
rs5803440	1	704	0	0	0	0						
rs5814559	1	700	0	0	0	0						
rs5820291	3	704	0	0	0	0						
rs5844947	1	696	0	0	0	0						
rs5848002	1	686	0	0	0	0						
rs5848302	1	696	0	0	0	0						
rs5883925	1	696	0	0	0	0						
rs5888463	1	702	0	0	0	0						
rs9332131	1	684	0	0	0	0						
rs9332736	28	700	4	0.006	0.120	1	rs4151672	0.045	95.4	4.0	0	0.6
rs9333357	1	704	0	0	0	0						
*TGFβR1**6A	9	3142	340	0.098	0.823	0.98	rs7850895	0.085	90.1	0.1	1.4	8.3
APC E1317Q	0	3472	30	0.009	0.051	0.914	rs10515449	0.12	87.5	11.7	0.1	0.8
rs28931588	0	634	0	0	0	0						
rs28931589	0	634	0	0	0	0						

Alleles M and m are for the overall cohort (cases and controls), tagSNPs denotes the SNP on the Hap550 arrays with the highest r^2 with the tested variants. Δ represents the number of bases inserted or deleted. The last four columns denote the haplotype frequencies, with the first letter being the allele given in columns 3 and 4, and the second letter being the major (M) or minor (m) allele of the tagSNP.

8.3 Discussion

The data presented in this chapter again underline the problems faced by a hypothesis driven marker search: unless the prior hypothesis is very compelling, verification is difficult. None of the markers analysed showed an association with CRC risk despite what at first glance appears to be a reasonable biological rationale. In over half the markers in the insertion/deletion screen, the minor allele was so uncommon that genotyping is unlikely to be worthwhile due to the large number of samples which would need to be typed to reach sufficient levels of power. It is conceivable that rs11509437 is associated with CRC, and will be typed in the whole Corgi set, with power of 80% to detect an association based on the observed allele frequencies. For two SNPs with a suggestion of an association (rs3069752 and rs3217458), the sample size analysed was too small to definitively exclude an association, but as they are well tagged by SNPs on the Illumina Hap500 array, r^2=0.99 and r^2=0.98, respectively, it is unlikely that typing either in the whole Corgi cohort would add to the knowledge derived from the analysis of the tagging SNPs.

The data in this chapter do show that tagSNPs can tag any locus that shows variability, but that the tagging is only efficient if the variant is sufficiently common. Taking the reverse approach to a GWAS, in this chapter putative CRC susceptibility loci were screened in the samples of the Corgi cohort and, as none of the variants were on the Illumina Hap550, the tagging efficiency determined retrospectively. The majority of the variants analysed were insertion-deletion changes, and the implication for GWAS in CRC, or any other disease, is that they will be able to detect non-SNP based variation in just the same manner as they can detect SNP based variation, thereby including an important pool of potential susceptibility loci. The flip-side is that rare insertion-deletion changes also behave like SNPs with low MAF and are thus not efficiently tagged. Therefore they are unlikely to be detected by GWAS with current SNP typing arrays. The variants tested in this chapter were not designed to find the lowest MAF for which tagging was efficient, but it appears that below about MAF=0.1, tagging as measured by r^2 will become inefficient using current SNP arrays. It may be that if the cut-off MAF for SNPs included on the high-density arrays were less than the current MAF≥0.05 (for the Illumina arrays used here), those rare, non-monomorphic variants could achieve higher tagging efficiency.

It is also clear that *D'* is not the appropriate measure to determine tagging efficiency of rare variants, as the nature of the *D'* measure yields a high value when the rare allele of a putative locus is almost invariably associated with one of the alleles of a nearby tagSNP, but that the reverse is not true, that the allele in question of the tagSNP does not reliably

inform about the status of the rare variant. If for the putative, rare variant the MAF is low, say MAF=0.01, and that for SNPs included on current arrays MAF≥0.05, even if the minor allele of the rare variant is always associated with the minor allele of the tagSNP, the reverse would be true in at most 20% of cases. Therefore, r^2 should be used as the measure of tagging efficiency for rare variants, and as the interest is in determining exactly the genotype at untyped loci even if the MAF is higher, r^2 should be the measure of choice for tagging efficiency for all loci.

Consequently, in order to capture rare variants, some degree of fine-mapping will be required. This could be performed subsequent to screening using 'conventional' SNP arrays, but it appears unlikely that a region harbouring a relatively highly penetrant but rare variant would be detected in the first instance[360]. It is possible that lower level significance across a region harbouring a rare variant could be detected, but at a lower level of significance, the noise-to-signal ratio would be high due to multiple testing producing falsely positive associations, which could also span across a region with high LD between the common variants typed. On the assumption that algorithms could be devised that would allow to gain some higher degree of certainty, it would still be likely that many regions showing such a weak association would need to be taken into the verification phase. This would have considerable resource implications, financially, manpower, and availability of DNA. As part of the work presented in this chapter, no attempt was made to determine the level of significance that might be expected across a region harbouring a true causal, rare variant, nor to devise any algorithms that might distinguish true low-level significance from false-positive findings.

Another approach would be to 'fine-map' using higher density SNP arrays, consisting not only of SNPs with MAF>0.05, as is the case for most of the non-CNV content of commercial arrays at present, but also SNPs with MAF≥0.01 or even less. Achieving sufficient density for such SNPs is likely to present considerable challenges, as determination of MAF in that order of magnitude requires the typing of hundreds to thousands of normal controls to reliably determine the allele frequencies, many more probands than have been typed as part of the HapMap project[348], the current selection basis for SNP arrays. Recognising the gap in the knowledge of genetic variation, those that lie between the very rare but highly penetrant loci, captured by traditional linkage analysis of affected families, and the common variant with low penetrance, captured by current SNP arrays, the 1000 genomes project was launched in 2008. It aims to generate very dense maps of human variation across many genetic backgrounds[361] to enable rare variants occupying the 'middle ground' to be examined for associations with disease. The implication is that the GWAS performed to date will have to be repeated, with new,

higher density arrays, more participants to generate the prerequisite power, and the attendant cost of performing such a study.

Despite the challenges facing the unbiased search for susceptibility loci for CRC, or any other complex trait, it still holds the greater promise for discovery of novel loci, and the hypothesis generation necessary for furthering understanding of the molecular biology of the cell, than some strands of current hypothesis driven research. None of the loci investigated in this chapter, irrespective of the strength of the prior hypothesis were shown to be associated with CRC risk, despite leveraging a large cohort of cases with a strong familial CRC background and hypernormal controls, underlining again the importance of verification of an initial hypothesis, a premise true for both hypothesis driven and unbiased marker discovery.

Chapter 9

Conclusions

The initial premise of this thesis was that the germline might be the driving factor for CRC risk, even if not all predisposing loci are discovered, and subsequently prognosis, in an as yet unquantified measure. Prognosis could be driven by factors that alter the likelihood of metastasis at an early time point, rendering surgery a non-curative intervention and increasing the role of adjuvant therapy, which may not be given because of the otherwise good prognostic pathological stage; or those factors that confer a growth advantage in the setting of hypoxia, chemotherapy, or to the host response of surrounding tissue. In the absence of any validated hypothesis of how the germline might influence prognosis, an unbiased, high-throughput screen of germline markers was adopted to try to address this question.

GWAS for prognosis are feasible, but the statistical analysis is crucial

The data presented in this thesis show that there is no intrinsic reason why GWAS aimed at teasing apart a complex genetic trait cannot be applied to outcome. Many of the necessary components are available to perform such a GWAS: large patient series with good quality follow-up and germline DNA, reliable, high-density SNP typing platforms, the statistical tools to analyse the data, and roadmaps to execute this GWAS successfully. A large cohort of patients with stage 2 and 3 CRC with follow-up data to clinical trial standard was identified, good quality DNA had been extracted, and this was analysed with excellent results on Illumina Hap300 or Hap370 BeadChips. Following the statistical analysis, the top candidate SNPs were typed in further verification cohorts, and the cohorts meta-analysed for patients with all stages of CRC and restricted to stage 2 and 3 patients. The most consistently significant SNP across these analyses was rs7556894, although it did not reach formal genome-wide statistical significance. For most SNPs analysed, restricting the analysis to patients with stage 2 and 3 disease, and for the top significant SNPs, adjusting for stage, increased the significance levels. Again, genome-

wide statistical significance was not reached. In an analysis to estimate the potential effect sizes of SNPs associated with prognosis, the estimates of the ES for the screening cohort were regressed to the mean, and it would appear that some SNPs could have HR≥1.5, an effect that, if proven, would have a major clinical impact. Based on that analysis, adjusted for stage, again rs7556894 was the most significant SNP, and on the assumption that ES and significance levels are already adjusted and correction is therefore only required for the multiple testing of the 39 SNPs included in verification, would have reached formal significance at a Bonferroni corrected $p<0.05$ ($p=7.25e-04$).

To date, the statistical tools had been applied only to the detection of genetic risk of a wide variety of malignant disease, e.g. colorectal[42,43], prostate[38], breast[39], ovarian[362], skin cancer[363], lymphoma[364], childhood ALL[365], melanoma[366], glioma[367]; non-malignant disease, e.g. cardiovascular disease[40,368], type 1 and 2 diabetes[41,369], multiple sclerosis[370]; and also for normal physiological variation such as age at menarche and menopause[263] or height[264]. In the larger studies, the risk replicated not only in the verification analyses by the same group, but also across different groups attempting to define novel risk loci at the same time. The most notable example of this regarding the susceptibility to CRC is rs6983267 on chromosome 8q24, which was published independently by two groups at the same time[44,299]. For some studies, one of the criticism is that the number of cases in the discovery phase was small, decreasing the power for detection[364,365]. For susceptibility, however, it is relatively easy to ascertain controls, and to a certain extent cases, as this is done on a snapshot-in-time basis, and power can thus be increased. For the GWAS presented here, and any GWAS for prognosis, this ascertainment is much more difficult, as all 'controls' still need to have the disease in question, but without relapse, and that both relapsed and non-relapsed patients need to have been followed up for at least three years, depending on the stage of disease investigated. This restricts the ability to conduct a GWAS for prognosis, and is generally cited, together with the relatively low event rate in the adjuvant setting, as the reason why none have yet entered the public domain.

A further draw-back of GWAS for prognosis is that in order to capture all survival information, a time-to-event analysis should be performed, unless the event rate is low, when tests of proportion such as the $\chi 2$ or Fisher's exact test give very similar estimates of significance to a time-to-event test[303]. In the discovery phase, the logrank test did not perform optimally in the recessive model due to its asymptotic nature and the poor estimate of the true effect in a small number of cases[302], and an over-representation of this genetic model among the SNPs taken forward from the discovery phase. This problem is not encountered in the GWAS of disease susceptibility as, generally, allele based tests (allelic odds ratios or the Cochran-Armitage test) are used. These tests also have their critics because, compared to using patients and genotype categories, there will be a

doubling of the numbers in each category (of alleles)[371], but appear to have been robust in the determination of many replicated susceptibility loci[44,52,299,372]. Ultimately, the approach is valid provided the locus in question is in HWE, and with allele counts, the more robust tests of proportionality can be used and the problem of small numbers in a category does not arise.

Therefore, the choice of survival analysis is critical for the success of a GWAS for prognosis. One such approach is presented in Chapter 5, where the data were analysed by genotype - analysis by allele does not allow time-to-event data capture - without further genetic models to avoid the over-representation of recessive model SNPs in the final analysis. The survival analysis was adjusted for stage, age and sex to avoid detecting SNPs that might be associated with stage rather than prognosis per se, and because the most significant SNP in the previous analysis had become more significant in multivariate analysis including stage. In addition, the discovery phase was extended to two cohorts with subsequent meta-analysis of all SNPs to increase power.

Many cohorts of sufficient size do not report DFS, but may have disease-specific overall survival, which is tightly correlated with DFS (r^2=0.90) in clinical trials evaluating interventions[315]. It is therefore reasonable to meta-analyse the two cohorts using a fixed effects model, provided that the statistical heterogeneity is not I^2>50%. In that case, it is likely that despite a marked effect in one cohort, the failure to replicate with a similar HR, i.e. without statistical heterogeneity, is suggestive of a spurious finding and these SNPs should not be taken forward into further verification.

Prognostic GWAS identify loci that largely fall into plausible functional categories

Despite the drawbacks of the GWAS outlined above, the top SNP identified appears promising. It almost reaches genome-wide statistical significance, is most significant in the analysis restricted to stage 2 and 3 CRC as would be expected having performed screening in that setting, and falls into a functional category that would be a plausible candidate for a prognostic marker. In trying to determine a realistic effect size conferred by this SNP, the Victor data was shrunk[310], but this also decreased the statistical significance. Using the shrunk values in stage 2 and 3 CRC, the ES for rs7556894 was HR=1.52, 95% CI 1.17-1.96, p=1.60e-03, increasing in significance to HR=1.56, 95% CI 1.21-2.03, p=7.25e-04 if this analysis was stage adjusted in multivariate analysis. It is likely that the effect size, due to the drawback of the time-to-event analysis with small numbers in the minor homozygote group, is more in keeping with that derived from the meta-analysis using the VICTOR data in its stage adjusted, shrunk form.

For the SNPs most consistently at the top of the list in terms of significance levels (rs7556894, rs672757, rs764372, rs4649314, rs472660), the function of the genes

associated with these loci appears largely plausible. rs7556894 is located on the short arm chromosome 2 at 2p14, 15Kb downstream from Actin-related protein 2 (*ARP2*), lying within an LD block covering the 3' end of the gene. ARP2 is a 339 amino acid protein and a member of the ARP2/3 complex involved in branching actin filaments. These are essential for cell shape and motility[307]. The ARP2/3 complex is increasingly expressed in CRC progression, both by tumour and stromal cells, and may thus provide a suitable environment for tumour invasion and metastasis[373]. Further, cancer cells require ARP2/3 to form actin-rich membrane protrusions (invadopodia) for the process of invasion[312]. Other elements required for the formation of invadopodia are cofilin and LIM kinase[290]. rs5997921 identified in the discovery phase, lies within LIM-kinase 2 on chromosome 22, and although this SNP did not replicate, it is attractive to speculate that germline variation affecting genes within the invadopodia pathway may be related to prognosis. rs4649314 is located on the long arm of chromosome 1 at 1q42.2, 50Kb downstream of mixed lineage kinase 4 (*MLK4*), a mitogen-activated protein kinase kinase kinase (MAP3K). MLK4 activates the JUN N-terminal kinase (JNK) and p38 mitogen activated protein kinase (MAPK) pathways[304] and these pathways are involved in a wide variety of cellular functions (e.g. proliferation, apoptosis, migration[305]) and alterations in the MAPK pathway have been implicated in colorectal carcinogenesis[306]. There is no evidence linking *MLK4* to CRC tumorigenesis, but further suggestive data for an involvement of the *MAPK* pathway in CRC prognosis comes from the gene 25kb telomeric of rs7556894, sprouty-related protein with EVH-1 domain 2 (*SPRED2*), which belongs to a family of proteins that regulate growth factor-induced activation of the MAP kinase cascade via their effect on the RAS/ERK pathway[374]. In addition, rs764372 lies in intron 5 of RAB31 on chromosome 18p, a member of the Ras superfamily of small transforming GTPases what ultimately converge onto the MAPK pathway. Again, no evidence for a role of *RAB31* in CRC exists, but over-expression of *RAB31* has been associated with poorer prognosis in lymph node negative breast cancer[375].

rs472660 is located intronic in *CYP3A43*, a cytochrome P450 detoxification gene that would be an interesting candidate, although more plausible as a predictive marker. It did not reach p<1e-04 in the screen for predictive markers for 5-FU, likely because *CYP* genes do not play a major role in the metabolism of 5-FU. It could be that the SNP tags *GJE1* (gap junction protein epsilon 1), another candidate from the cell adhesion functional family, tying it into the same functional category as rs7556894. The SNP for which there is no obvious potential link to prognosis is rs672757 in intron 2 of *KCND3*, a K^+- voltage-gated channel on chromosome 1. While there has been no link to cancer, *KCND3* is associated with small G-protein Rap signalling in response to angiotensin 1 (ref [376]); like rs472660, it would be a better candidate as a predicative marker.

The concept of prognosis being linked to early metastatic potential, and thus the promise to be able to better predict who will have micrometastatic disease and offer adjuvant chemotherapy is attractive. It would also be relevant in stage 3 CRC, where patients have pathologically proven lymphnode metastasis, yet a large minority appear cured with surgery alone. In this setting, it would not be the offer of adjuvant chemotherapy that changed, but rather the type of regimen, length of treatment, and follow-up procedures to manage a potential relapse early, although in the absence of good prognostic markers beyond pathological stage, none of these avenues have been developed or shown to be more effective compared to the current standard of care.

Another strand of suggestive evidence for the concept of a genetic locus being associated with early metastases that may not have been detected pathologically, comes from the findings presented in Chapter 6, where rs6983267 appears to be associated with micro-metastasis and an at least borderline significant association with survival.

A functional category not represented in the top 40 SNPs from the discovery phase was SNPs that reside in or are close to genes known to be involved in the metabolism of 5-FU, the most obvious functional area where germline variation might play a role, and the only area where there is at least a weak body of evidence suggesting an association with prognosis[206,211,222]. This is to be expected because the discovery phase did not take into account the administration of adjuvant chemotherapy, and this information was only available for one of the verification cohort, PETACC , in which all patients received 5-FU, but genotype data, compared to VICTOR, was very restricted. Therefore, no formal verification of pharmacogenetic markers was undertaken, which might have yielded predictive, rather than prognostic, markers. The two SNPs for which there was data available from both cohorts did not reach genome-wide significance, and the restriction to patient who had received adjuvant chemotherapy also does not guarantee that the marker is predictive as such an analysis could also yield prognostic markers. Despite this, there is potential to perform such an analysis in patients who had received adjuvant chemotherapy. For example, the SNP from the discovery phase that is most plausible to carry predictive information, rs472660 in *CYP3A43*, also had a large effect going in the same direction in VICTOR and PETACC when restricted to patients who had received adjuvant chemotherapy, reaching significance at the $p<0.05$ level in the latter cohort (Table 4-2). The meta-analysis of just those two cohorts gave a p-value that almost reached genome-wide significance levels (p=6.04e-07, data not shown). This raises the possibility that the causative locus tagged by the minor allele of rs472660 may be a novel determinant of the usefulness of 5-FU in the adjuvant setting.

Further risk markers are being discovered, with diminishing rates of return

Following the initial successes in the detection of CRC susceptibility loci using unbiased, high-throughput GWAS, the discovery rate is diminishing, with larger and larger patient cohorts being required to detect an association. rs6687758, evaluated in Chapter 7, represents such a locus, reaching genome-wide significance at p=5.04e-08. In fact, rs6687758 had not been further evaluated in the initial meta-analysis of the four cohorts, which had been powered to detect common loci conferring quite large risks, generally larger that found for rs6687758 (OR=1.15). It is unlikely that there are many further CRC susceptibility SNPs with OR≥1.2 and minor alleles with frequencies >0.2 in populations of European ancestry. This underlines the efforts needed to detect low-penetrance variants that have OR<1.2 and/or have MAF<0.2, as variants with these characteristics are likely to make up a further proportion of susceptibility loci for CRC. This implies expanding the scale of GWAS by including more SNPs in the initial screen, increasing sample size and the number of SNPs taken into the replication phase, with all the costs that this would involve. With the diminishing returns of the current strategy[377], however, other sources of genomic variation will need to be considered, and in particular CNV[378]. There is now an increasing appreciation of the extent of CNV throughout the genome[379-383], but the application of GWAS techniques to CNV has been limited so far. This has been driven by the poor coverage of CNV regions in earlier arrays[384], and the signal-to-noise ratio for CNV inferred from intensity data at each analysed locus, which can be as low as 1.2 for some arrays[385]. Both are now being addressed by the major SNP array manufacturers[345], as advances in SNP typing technology and CNV calling algorithms[386] also improve the signal-to-noise ratio, and the first studies of genome-wide CNV analysis for cancer susceptibility have now been published[387].

The concern that GWAS might nonetheless 'discover' spurious associations remains, as not all described loci across the various diseases have an obvious functional consequence, or are even in or near a gene with a suggestive functional relevance. It is unlikely that all published susceptibility loci will have their functional consequences elucidated in the near future, but inroads are being made to address the lack of an immediate functional explanation of risk. rs6983267, located on chromosome 8q24, lies intergenic but nearest to *POU5F1P1*, a transcription factor of undetermined functional significance, and about 337Kb from *MYC*, a target of *WNT* signalling. *MYC* acts downstream of activated β-catenin in the *WNT* pathway, which jointly with TCF4 enhances *MYC* expression via the TCF4 binding site in the *MYC* promotor[336]. rs6983267 also lies in a T-cell factor (TCF) consensus binding sequence, and the two alleles show differential binding of TCF4[337] and TCF7L2[332], a further effector of *WNT* signalling[388] and carrying itself a SNP associated with CRC risk[389,390]. This region physically interacts with the *MYC* promotor, although an

effect on *MYC* expression could not yet be established[332,337]. Fine mapping of the 8q24 region did not reveal any further pathogenic variants, and it appears likely that rs6983267 is indeed the causal variant[391].

SNP arrays can tag any genetic variation, but may not capture it well

In the search for new determinants of complex traits, the concept of assessing the whole genome for genetic variation by only typing a relatively small number tagSNPs is very attractive as it reduces cost significantly. This approach has been proven useful by the many examples of GWAS mentioned above. The tagSNP itself can be the determining genetic alteration, as appears to be the case for rs6983267, but most tagSNPs are expected to be only markers for a causative variant. Therefore, it is important to understand the efficiency with which the tagSNP provides information about the variation within the tagged LD block. Every SNP typing platform contains SNPs that tag other loci, only the strategy for choosing the SNP content varies: for the first wave of SNP arrays, Illumina adopted an approach where the SNPs included in the array are those HapMap SNPs that performed best within a given LD block defined by *D'*, while other array providers focussed more on gene-centric SNPs.

The premise of the Illumina BeadChips was that they were designed to capture the majority of common SNPs with MAF>0.2 in European populations with $r^2 > 0.80$. With the initial limitation of the number of SNPs included in the array, only around 10% of SNPs with MAF=0.05-0.1 are tagged at $r^2 > 0.80$, thus limiting the ability to tag this type of SNP[360]. Further, CNV was not tagged efficiently, because SNPs in CNV high regions do not perform well in the general measures of assay quality[392], leading to bias against them in the selection of array content and subsequently poor coverage of CNV regions[384]. Lastly, it was not clear if insertion/deletion polymorphism would behave in the same fashion as ordinary SNPs, and what the tagging efficiency would be.

Based on the data presented in Chapter 8, the insertion/deletion variants appear to behave like ordinary SNPs, but with minor allele frequencies well below the common variant 'threshold'. Like other rare variants, they are insufficiently captured by current SNP arrays, but at least it appears that the question of tagging any particular locus is reduced to MAF frequency and copy-number, and the response to this the improved physical coverage of the genome, and its regions of interest, and the MAF of the tagSNP to efficiently tag CNV and rare variants, respectively.

The biggest step forward has come from the inclusion of SNP or monomorphic probe content aimed specifically at CNV regions, and increasing the number included on the arrays to achieve greater physical coverage of the whole genome including CNV. With the data generated from these, CNV analyses become more robust, but still have the

caveat of an inferior signal-to-noise ratio compared to the genotyping performed at the same time. The higher density of SNPs, and inclusion of rarer SNPs begins to address the issues of rarer variants, but the knowledge of such rare variants remains limited due to the small number of individuals typed for the HapMap SNP database[348], which at present forms the overwhelming basis for tagSNP selection. Inclusion of SNPs identified in the 1000 Genomes project[361] will help, but the tagging efficiency for SNP selection should not be based on *D'* but rather on r^2, as many rare variants already have *D'*=1, despite the inability to predict the minor allele reliably (see also section 8.3).

It is therefore plausible that there are many other loci with similar genetic characteristics to rs6687758 that remain to be discovered, where the power of current GWAS is too small to detect other, similar loci. At the same time, the population attributable risk of such loci is likely to be small[393]. The future may therefore be the search for variants that are rare enough not to be captured by the common-variant GWAS, nor penetrant enough to cause familial clustering and be detected by traditional linkage analysis[394]. The number of loci to be discovered that way, however, could be small, and , for example in type 2 diabetes, on the assumption of MAF≤0.01 and OR=3, the remaining inherited risk could be accounted for by 30 loci[393]. Increasing SNP array density and improving our knowledge of rare variants will help this cause, but to have the power to detect such effects given the low MAF will drive up size, cost and effort. How such arrays would deal with rare CNV is not clear, and the possible effect of epigenetic risk factors would only be captured imperfectly as epigenetic changes can depend on underlying DNA sequence[395], but may not always do so[396]. It is possible that next-generation sequencing (NGS), i.e. highly parallel, high-throughput sequencing[397], may hold the key to these questions, although the statistical tools to take full advantage of the vast amount of data generated by NGS are still being developed[398].

The germline as the overarching paradigm? Yes and No

The germline clearly carries the blueprint for many diseases, and regarding susceptibility to CRC, it clearly contributes to a large proportion of cases; loci driving this process have been and are being identified and evaluated. The common variant-common disease hypothesis has advanced our knowledge of CRC risk, without fully explaining the inherited proportion. Future GWAS will add to this knowledge, and if the lessons from earlier GWAS are learnt, particularly regarding rare variants, then a large part of familial CRC risk is likely to be explained.

For prognosis, the jury remains out, and the data presented here do not preclude an effect without having been able to prove an association. Several of the promising candidates will be typed in further cohorts to derive a final assessment of these. Improving power and

statistical analysis will help, should the former approach ultimately fail, but again, like with GWAS for disease susceptibility, further verification to drive precision and power of the final estimate will be imperative. At present, the only genetic data for an effect on prognosis in CRC remains the association of resistance to anti-EGFR strategies with *KRAS* mutations. It is likely that somatic changes hold the majority of prognostic and predictive information, but it is conceivable that the somatic changes may be driven by germline variation. An example of the germline driving somatic changes would be HNPCC, but there is no clear data to suggest that HNPCC patients have the same, better prognosis that is conferred by sporadic MSI+ CRC[188]. It is also conceivable that the germline does not carry prognostic information with any clinical utility: several GWAS for disease susceptibility in haematological malignancies had a similar size and case-control ratio to the work in this thesis, which covered 700 cases and 2146 controls with stage 2 and 3 CRC, yet these studies did identify novel disease susceptibility loci[364,365,399]. The work in this thesis is therefore likely to have excluded a very strong (HR≥2) effect of germline SNPs on prognosis.

Further work is required before any germline locus can be thought of as being associated with prognosis or the response to therapy in CRC, and be incorporated into clinical decision-making algorithms. In addition to the work presented in Chapters 3, 4, and 5, an increasing understanding of tumour biology will lead to evaluation of more hypothesis driven markers, just as unbiased screening will generate more hypotheses to be tested. However, even for better validated markers, there is often little consensus about their value: the prognostic impact of the molecular phenotype has been known about for many years without driving prospective studies or influencing clinical decisions. It is clear that a much greater degree of cooperation is required between basic and clinical scientists to bring about trials with enough statistical power to provide results that will compel clinical engagement.

Appendices

Appendix A: Tables

Table 10-1 URLs to resources used in this thesis

Utility	Source
CaTS	http://www.sph.umich.edu/csg/abecasis/cats/reference.html
dbSNP	http://www.ncbi.nlm.nih.gov/SNP/
Eigensoft	http://genepath.med.harvard.edu/~reich/Software.htm
Haploview	http://broad.mit.edu/mpg/haploview
NCBI UniGene	http://www.ncbi.nlm.nih.gov/UniGene/
Primer3	http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi
UCSC Genome Browser	http://genome.ucsc.edu/

Table 10-2 Functional annotation of SNPs chosen only on p-value

SNP	Description
rs7556894	15Kb downstream from Actin-related protein 2 (*ARP2*), a member of the ARP2/3 complex involved in branching actin filaments and essential in cell shape and motility[307]. It lies within the LD block covering the gene based on *D'*, but not r^2. A little further away lie sprouty-related protein with EVH-1 domain 2, member of a family of proteins that regulate growth factor-induced activation of the MAP kinase cascade (25kb), and Ras-related protein Rab-1A (155kb).
rs2589183	In intron 1 of syndecan-2 precursor, a transmembrane heparan sulfate proteoglycan, and may influence tumour behaviour through regulation of adhesion or proliferation in CRC[400]. rs2589183 lies between the LD blocks of the exons either side, and there are no further suggestive SNPs in the region on Manhattan plot.
rs764372	In intron 5 of *RAB31* on chromosome 18p, a member of the Ras superfamily of small transforming GTPases. It lies at the heart of a weak LD block in intron 5. The over-expression of *RAB31* has been associated with poorer prognosis in lymph node negative breast cancer[375].
rs672757	In intron 2 of *KCND3*, a K^+- voltage-gated channel on chromosome 1. While there has been no link to cancer, *KCND3* is associated with small G-protein Rap signalling in response to angiotensin 1 (ref [376]). Other SNPs which are part of the same LD block were also weakly associated with prognosis in the discovery phase.
rs7912136	128Kb upstream of protein kinase C, isoform θ on chromosome 10. Protein kinases have been implicated in many cellular processes, there is no specific evidence linking PKC, isoform θ to cancer. Of note, however, is that there are several SNPs with somewhat lower levels of significance located within protein kinase C, isoform θ, which are in long-range LD with rs7912136.
rs6972789	350Kb downstream of stromal antigen 3-like 4 on chromosome 7 as part of an LD block extending towards but not reaching stromal antigen 3-like 4. The annotation of the UCSC genome browser suggests that stromal antigen 3-like 4 is down-regulated in CRC. Other SNPs within that LD block also show weak evidence for an association with outcome on the Manhattan plot.
rs7007146	108Kb upstream of E3 ubiquitin protein ligase (also known as *EDD*) on chromosome 8q22, a region gained in many breast cancer, where this has been linked to poorer prognosis. However, it appears that metadherin may be the causative gene[401], located 5Mb from rs7007146. Over-expression of *EDD* has been associated with worse-prognosis and resistance to cisplatin therapy in ovarian cancer[402].
rs745888	Located just downstream of complement component 1-like (*C1QL2*) on chromosome 2. It is part of an LD block extending to the 5' end of *C1QL2*, and part f a weak LD block extending to dudulin 2 (*STEAP3*). Dudulin 2 is differentially expressed in response to TP53 (ref [403]) and could be a regulator of apoptosis[404]. Further, there is a SNP with weak evidence located downstream of dudulin 2.
rs4978394	Located just upstream of protein tyrosine phosphatase N3 (*PTPN3*) on chromosome 9, part of a weak 200Kb LD block upstream of *PTPN3*. rs4978394, together with rs10816807 - itself associated with outcome at the 1e-04 level and included in the list of 109 SNPs under consideration for verification - flanks PTPN3 either side. *PTPN3* has been shown to be mutated in a subset of CRC[405].
rs3784780	In intron 9 of Bloom's helicase on chromosome 15. Bloom's helicase is a DNA helicase with a probable role in DNA replication and double-strand break repair, and the gene underlying Bloom's syndrome, an autosomal recessive disorder characterised by frequent chromosome breaks and a predisposition to many cancers. The SNP lies within the LD block covering most of the 5' half of the gene, although there are no other suggestive SNPs in the region. Of note is that the strongest association with survival is in the genotypic model and it appears that the heterozygote carriers have the best prognosis, while homozygote AA carriers have the worst. The resulting p-trend$_{gen}$=0.417.

Table 10-2 Functional annotation of SNPs chosen only on p-value (cont'd)

SNP	Description
rs10510044	130Kb downstream of the receptor for prolactin releasing hormone (*PRLHR*), a G-coupled receptor largely found in the anterior pituitary. It is located in a large LD block, extending away from *PRLHR*, which also contains further SNPs with weak evidence for an association with survival. There is no evidence linking *PRLHR* to cancer.
rs712082	In expressed sequence tag (EST BU947851) on chromosome 1, in a 35Kb block of high LD which includes two further SNPs associated with prognosis (rs698277 and rs712068) with virtually identical genotype distribution (and therefore not included in phase 2). rs712082 lies 80Kb upstream of cornichon homolog 3, a family of proteins involved in the transport, processing and secretion of TGF_{α}, a known ligand of the epidermal growth factor receptor[406].
rs11788150	250Kb downstream of nuclear factor I/B (*NFIB*) on chromosome 9, in a weak LD block extending almost to the 5' end of *NFIB*. Nuclear factor IB has been described as a translocation partner in colonic lipomas[407], but while it has not been linked to adenocarcinoma, *NFIB* is down-regulated in colon cancer cells (http://symatlas.gnf.org/SymAtlas/). There is a further SNP in this LD block which reached the cut-off of 1e-04 and was included in the selection in section 3.6.2 as they are Only in moderate LD (r^2=0.52) even if *D'* is higher (*D'*=0.92).
rs10842099	Located within Homo sapiens cDNA FLJ37414 as a solitary 'hit', with no other genes in the vicinity.
rs4715476	370Kb downstream of the tubulointerstitial nephritis antigen (*TINAG*), a novel basement membrane protein interacting with laminin and type IV collagen promoting cell adhesion[408]. There is a weak LD block extending to the 5' end of *TINAG*. There is a hypothetical gene 80Kb centromeric to rs4715476 (*LOC222584*), with a strong LD covering this locus and rs4715476. There is no evidence for any other SNPs in the region.
rs1526884	Located on chromosome 19.12, among a cluster of zinc-finger proteins (ZNF). It lies between *ZNF257* and *ZNF676*, downstream of both. There is high LD in this region, with the LD block in which rs1526884 lies extending towards and beyond *ZNF257* with *D'*=1.0. r^2 is much lower, and only with a cluster of SNPs immediately upstream of *ZNF257* (rs431346 to rs408312) is it relatively high at r^2=0.51. A further LD block, of which rs1526884 is not part, extends towards ZNF676, but rs1526884 has high r^2 with a SNP immediately downstream of *ZNF676* (r^2=0.87), and high *D'* with the whole gene.
rs4649314	50Kb downstream of mixed lineage kinase 4 (*MLK4*), a mitogen-activated protein kinase kinase kinase (MAP3K) on chromosome 1, and known activator of the JUN N-terminal kinase (*JNK*) and p38 mitogen activated protein kinase (*MAPK*) pathways[304]. The interplay of these pathways is involved in a plethora of cellular homeostatic functions (e.g. proliferation, apoptosis, migration[305], and alterations in the *MAPK* pathway have been implicated in colorectal carcinogenesis[306], although there is no evidence that *MLK4* specifically is a factor in CRC. There is an LD block extending towards *MLK4*, with the region from exon 7 to 10 having a high *D'* with rs4649317 (*D'*=0.69-1.0), but low r^2 with r^2=0.06-0.42. Of the SNPs on the arrays covering this region, no other was significantly associated with prognosis.
rs1350308	In intron 1 of CUB and Sushi multiple domains 1 (*CSMD1*) on chromosome 8q23. *CSMD1* is a huge gene with 3565 amino acid protein spanning over 2Mb and is of unknown function. *CSMD1* covers a minimal region of deletion associated with higher stage in prostate cancer, and it has been linked to higher stage in CRC[297]. The LD structure of the region displays weak LD blocks, and there is no evidence of long-range LD covering the exons flanking rs1350308. This is supported by the Manhattan plot, which shows that in the immediate vicinity, there is some weak evidence for an association with outcome.

Table 10-2 Functional annotation of SNPs chosen only on p-value (cont'd)

SNP	Description
rs9533457	In intron 9 of ecto-NOX disulfide-thiol exchanger 1 (*ENOX1*) on chromosome 13. *ENOX1* is a plasma membrane bound electron exchange protein[409], as part of the plasma membrane redox system, implicated in control of cell growth and proliferation, among others[410]. rs9533457 lies within an LD block covering exons 9 and 10.
rs1571583	In GLIS family zinc finger 3 isoform b, a Krüppel like zinc finger transcription factor that has been linked to neonatal diabetes. It has not been linked to CRC, but acts synergistically with BMP2 on osteoclast differentiation[411], and in turn BMP2 has been linked to CRC risk[46].
rs567564	Located 105Kb from FRK tyrosine kinase, also known as *RAK*, interacts with the retinoblastoma protein, with growth suppressing properties[412]. It may also have anti-proliferative properties in its own right[413], but has not been linked to CRC.
rs7866165	273b from Nuclear factor I/B, see entry for rs11788150.
rs437171	Not located near any gene, the nearest gene is Cadherin 18, type 2 1.3Mb away.
rs3752261	10Kb from Zinc finger protein 334 isoform b (*ZNF334*) belonging to the Krüppel C2H2 family of transcription factor without a known function. Members of this family have been found to interact with *MYC*[414].
rs1924597	*HS6ST3* modifies heparan sulfate so it can interact with a variety of proteins which in turn have been implicated in proliferation, differentiation, adhesion, and migration[415].

Among the initial 25 SNPs chosen on p-value alone, it would be expected that not all would present a plausible biological rationale for an association with prognosis. There were nonetheless a number of interesting candidate genes. The functional annotation of the 40 SNPs taken forward are listed in Table 3-7, are described in section 3.6.2 for SNPs chosen on biological rationale, and in more detail for those SNPs chosen on statistical significance below. For a distribution across functional categories, see Figure 3-16, Manhattan plots for all SNPs are given in Appendix B.

Table 10-3 List of 103 SNPs meeting all inclusion criteria

SNP	Chr	Base	Model	ES (95% CI)	p-value	Nearest gene
rs7556894	2	65.37	recessive	3.09 (2.00-4.77)	9.19e-08	Actin-related protein 2 (15kb)
rs2589183	8	97.59	allelic	0.37 (0.25-0.55)	1.06e-07	Syndecan 2 precursor
rs764372	18	9.82	recessive	4.24 (2.29-7.84)	5.09e-07	*RAB31*
rs672757	1	112.15	allelic	1.84 (1.44-2.36)	1.08e-06	KCND3 potassium channel
rs7912136	10	6.79	recessive	2.58 (1.73-3.85)	1.27e-06	Protein kinase C, theta (128kb)
rs6972789	7	66.76	recessive	2.17 (1.57-3.00)	1.42e-06	Stromal antigen 3-like 4 (350kb)
rs7007146	8	103.60	recessive	2.14 (1.55-2.95)	2.33e-06	E3 ubiquitin protein ligase (108kb)
rs745888	2	119.63	recessive	2.68 (1.75-4.12)	2.48e-06	Complement component 1 (4kb)
rs4978394	9	111.31	genotypic	1.72 (1.32-2.24)	2.48e-06	*PTPN3* (6kb)
rs3784780	15	89.11	genotypic	0.88 (0.64-1.21)	2.94e-06	Bloom syndrome protein
rs10510044	10	120.23	recessive	2.46 (1.66-3.64)	3.60e-06	*PRLHR* (130kb)
rs712082	1	222.79	recessive	3.01 (1.84-4.92)	3.89e-06	EST BU947851
rs11788150	9	13.82	recessive	2.32 (1.59-3.37)	6.77e-06	Nuclear factor I/B (250kb)
rs10842099	12	23.08	allelic	0.51 (0.37-0.69)	7.26e-06	cDNA FLJ37414
rs4715476	6	54.73	genotypic	1.73 (1.33-2.24)	8.12e-06	*TINAG* (370kb)
rs1526884	19	22.10	dominant	0.47 (0.33-0.66)	8.47e-06	near cluster of ZNF (31 - 750kb)
rs4649314	1	231.64	recessive	2.30 (1.58-3.35)	8.54e-06	Mixed lineage kinase 4 (50kb)
rs1350308	8	4.60	allelic	1.74 (1.36-2.22)	9.39e-06	*CSMD1*
rs9533457	13	42.81	recessive	2.58 (1.67-3.98)	9.62e-06	NOX disulfide-thiol exchanger
rs1571583	9	4.26	recessive	2.80 (1.73-4.53)	1.10e-05	GLIS family zinc finger 3 isoform b
rs567564	6	116.26	recessive	2.37 (1.59-3.53)	1.21e-05	*FRK*
rs7866165	9	13.80	recessive	2.85 (1.74-4.66)	1.23e-05	Nuclear factor I/B (273kb)
rs437171	5	18.17	recessive	3.08 (1.81-5.25)	1.28e-05	Cadherin 18, type 2 (1.3MB)
rs3752261	20	44.55	recessive	3.41 (1.89-6.15)	1.41e-05	Zinc finger protein 334 (10kb)
rs1924597	13	96.24	allelic	0.38 (0.23-0.61)	1.46e-05	*HS6ST3*
rs6487867	12	30.01	recessive	2.30 (1.56-3.39)	1.54e-05	Transmembrane / tetratricopeptide repeat (220kb)
rs1822917	8	72.43	recessive	2.65 (1.67-4.20)	1.55e-05	Eyes absent 1 isoform c
rs1411684	9	111.36	genotypic	1.69 (1.30-2.19)	1.56e-05	PTPN3 (100kb)
rs9514816	13	107.58	allelic	2.07 (1.51-2.83)	1.68e-05	DNA ligase IV (77kb)
rs6764062	3	152.85	recessive	3.37 (1.87-6.08)	1.76e-05	Arylacetamide deacetylase (90kb)
rs698277	1	222.79	recessive	2.86 (1.73-4.73)	1.91e-05	Cornichon homolog 3 (80kb)
rs1665853	4	181.29	recessive	3.00 (1.76-5.11)	2.09e-05	nothing
rs4411217	10	120.21	allelic	1.72 (1.35-2.21)	2.14e-05	PRLHR (130kb)
rs4771704	13	110.09	allelic	1.99 (1.47-2.72)	2.18e-05	Cysteinyl-tRNA synthetase 2
rs6518956	22	34.27	genotypic	0.97 (0.78-1.21)	2.37e-05	RASD family, member 2
rs6853554	4	107.17	allelic	0.53 (0.39-0.72)	2.42e-05	Nephronectin (54kb)
rs663472	12	3.97	recessive	3.18 (1.80-5.61)	2.46e-05	ADP-ribose polymerase (116kb)
rs1878632	2	153.03	recessive	2.06 (1.46-2.91)	2.58e-05	Formin-like 2
rs7221974	17	1.58	recessive	2.47 (1.60-3.81)	2.62e-05	Alpha-2-plasmin inhibitor
rs1465365	3	188.56	genotypic	0.94 (0.70-1.27)	2.80e-05	INF responsive protein (13kb)
rs7600624	2	152.99	allelic	1.96 (1.45-2.66)	2.91e-05	Formin-like 2
rs935361	2	119.65	recessive	2.94 (1.73-5.01)	3.05e-05	Complement component 1 (20kb)
rs9315425	13	35.98	recessive	3.58 (1.89-6.80)	3.09e-05	Cyclin A1 (65kb)
rs958882	5	123.03	recessive	2.45 (1.58-3.78)	3.15e-05	Casein kinase 1(50kb)
rs2827250	21	22.43	dominant	1.97 (1.42-2.73)	3.72e-05	Neural cell adhesion (604kb)

Chr - chromosome, base - position on chromosome in Mb.

Table 10-3 List of 103 SNPs meeting all inclusion criteria (cont'd)

SNP	Chr	Base	Model	ES (95% CI)	p-value	Nearest gene
rs968757	4	120.98	recessive	1.99 (1.42-2.77)	3.79e-05	Phosphodiesterase 5A (210kb)
rs4741450	9	15.07	recessive	1.96 (1.41-2.72)	3.87e-05	cDNA FLJ46077
rs4813685	20	4.24	recessive	3.00 (1.73-5.19)	3.92e-05	Alpha-1D-adrenergic receptor
rs3124028	9	7.66	allelic	0.59 (0.46-0.76)	4.05e-05	PTP (645kb)
rs11682848	2	234.58	recessive	3.35 (1.81-6.19)	4.18e-05	Transient receptor potential channel
rs1526643	2	76.34	genotypic	1.68 (1.26-2.22)	4.38e-05	*GFHL3075* (480kb)
rs2288742	16	25.03	allelic	0.59 (0.46-0.76)	4.57e-05	Leucine-carboxyl methyltransferase 1 (5kb)
rs712068	1	222.78	recessive	2.73 (1.65-4.52)	4.58e-05	Cornichon homolog 3 (80kb)
rs12587148	14	33.06	genotypic	1.10 (0.89-1.38)	4.75e-05	Neuronal PAS domain protein 3
rs10429965	1	209.91	recessive	3.68 (1.88-7.22)	4.78e-05	NEK2
rs5997921	22	29.96	recessive	3.68 (1.87-7.21)	4.88e-05	LIM domain kinase 2
rs10513640	3	22.21	recessive	3.67 (1.87-7.20)	4.96e-05	Zinc finger protein 659 (240kb)
rs6057021	20	9.85	recessive	2.36 (1.54-3.62)	4.99e-05	p21-activated kinase 7 (80kb)
rs1824476	18	27.28	allelic	0.55 (0.41-0.74)	5.10e-05	Desmoglein 3 (0.6kb)
rs4463351	7	99.92	recessive	3.04 (1.72-5.37)	5.28e-05	*TSC22D4* (1.5kb)
rs2291270	5	169.08	dominant	0.53 (0.39-0.72)	5.28e-05	*DOCK2*
rs7974465	12	75.65	genotypic	1.14 (0.91-1.43)	5.46e-05	Oxysterol-binding protein (168kb)
rs4776494	15	62.50	genotypic	1.77 (1.35-2.32)	5.66e-05	Thyroxine receptor interactor
rs6782375	3	105.16	dominant	0.53 (0.39-0.73)	5.81e-05	cDNA clone 6614812
rs4856091	3	105.12	dominant	0.53 (0.39-0.73)	6.01e-05	cDNA clone 6614812 (5kb)
rs221259	7	157.03	recessive	3.43 (1.81-6.51)	6.04e-05	PTP receptor type, N
rs2217375	8	112.63	recessive	3.43 (1.81-6.51)	6.04e-05	*CSMD3* (900kb)
rs1075737	7	102.86	recessive	3.43 (1.81-6.52)	6.06e-05	*SLC26A5*
rs657224	1	76.94	dominant	1.89 (1.38-2.60)	6.13e-05	*ST6GALNAC3* (115kb);
rs1366978	1	198.05	recessive	1.99 (1.41-2.80)	6.14e-05	Orphan nuclear receptor NR5A2
rs10800652	1	198.05	recessive	1.99 (1.41-2.80)	6.14e-05	Orphan nuclear receptor NR5A2
rs10876477	12	52.34	dominant	1.89 (1.38-2.59)	6.15e-05	ATP synthase (6kb)
rs1038923	12	65.14	allelic	0.57 (0.43-0.75)	6.20e-05	Glutamate receptor interacting 1
rs305002	15	67.93	dominant	0.53 (0.39-0.73)	6.25e-05	Transducin-like enhancer (211kb)
rs1928661	9	29.30	allelic	0.59 (0.45-0.77)	6.31e-05	Leucine rich repeat neuronal (599kb)
rs2280117	22	18.69	dominant	0.53 (0.38-0.73)	6.63e-05	DGCR6L
rs4553248	1	105.98	recessive	3.41 (1.79-6.47)	6.77e-05	cDNA clone 5297905 (17kb)
rs4242025	5	51.15	recessive	2.39 (1.53-3.72)	7.09e-05	*ISL1* (425kb)
rs10192011	2	53.66	recessive	2.81 (1.65-4.79)	7.17e-05	Ankyrin repeat/SOCS box-containing 3 (90kb)
rs215410	4	23.03	recessive	2.15 (1.46-3.18)	7.46e-05	*PPARGC1A* (370kb);
rs457918	20	44.55	recessive	3.23 (1.75-5.96)	7.50e-05	Zinc finger protein 334 (20kb)
rs10916777	1	20.58	dominant	1.87 (1.36-2.55)	7.53e-05	cDNA clone IMAGE:5269505
rs10740509	10	54.14	recessive	2.28 (1.50-3.48)	7.73e-05	MBL2 (50kb)
rs10816807	9	111.21	recessive	2.65 (1.60-4.39)	7.97e-05	*PTPN3*
rs10414779	19	57.49	recessive	2.25 (1.49-3.42)	8.06e-05	Zinc finger protein 766

Chr - chromosome, base - position on chromosome in Mb.

Table 10-3 List of 103 SNPs meeting all inclusion criteria (cont'd)

SNP	Chr	Base	Model	ES (95% CI)	p-value	Nearest gene
rs2327672	6	136.27	allelic	0.43 (0.28-0.68)	8.17e-05	Phosphodiesterase 7B
rs4812682	20	41.22	recessive	2.23 (1.48-3.36)	8.20e-05	Protein tyrosine phosphatase
rs1438620	2	222.75	dominant	2.25 (1.49-3.41)	8.25e-05	*PAX3* (26kb)
rs1094039	9	6.69	genotypic	0.96 (0.77-1.21)	8.30e-05	Glycine dehydrogenase (50kb)
rs1565741	8	16.80	genotypic	1.75 (1.35-2.27)	8.38e-05	Fibroblast growth factor 20 (95kb)
rs7974869	12	106.55	recessive	2.53 (1.56-4.08)	8.74e-05	BTB (POZ) domain containing
rs2807602	9	125.93	genotypic	1.67 (1.30-2.14)	8.77e-05	LIM homeobox protein 2 (93kb)
rs2760555	20	16.07	genotypic	1.64 (1.25-2.15)	9.35e-05	Kinesin-like motor protein (130kb)
rs7616534	3	198.78	recessive	3.51 (1.79-6.89)	9.46e-05	3-hydroxybutyrate dehydrogenase
rs2430893	18	52.58	genotypic	1.27 (1.02-1.59)	9.73e-05	*WDR7*
rs2514841	8	109.13	recessive	0.34 (0.19-0.60)	9.83e-05	*RSPO2*
rs10273192	7	149.57	recessive	2.31 (1.50-3.58)	9.90e-05	Actin-related protein 3-beta (7kb)
rs838308	12	113.43	allelic	0.51 (0.36-0.73)	9.96e-05	T-box 5 isoform 1 (100kb)
rs12330780	3	187.22	recessive	1.89 (1.37-2.63)	9.96e-05	Ets variant gene 5 (22kb)
rs1034116	14	89.02	allelic	1.89 (1.38-2.57)	1.14e-04	Checkpoint suppressor 1
rs25689	9	73.55	allelic	1.70 (1.29-2.22)	1.81e-04	Transmembrane protein 2
rs12583114	13	97.56	genotypic	1.58 (1.26-1.98)	2.15e-04	FERM, RhoGEF, pleckstrin protein
rs304998	15	67.93	allelic	1.61 (1.26-2.07)	2.43e-04	Transducin-like enhancer 3 (200kb)

Chr - chromosome, base - position on chromosome in Mb.

Table 10-4 Meta-analysis for all stages, adjusted for stage

SNP	Cohorts	Model	p-value	ES (95% CI)	I^2	p_Q
rs7556894	4	recessive	1.91e-05	1.60 (1.29-1.98)	75.8	6.14e-03
rs472660	3	recessive	2.63e-05	2.29 (1.56-3.37)	68.4	4.23e-02
rs7912136	2	recessive	3.47e-04	1.67 (1.26-2.22)	86.2	7.00e-03
rs764372	3	recessive	4.97e-04	1.92 (1.33-2.76)	79.0	8.64e-03
rs1878632	3	recessive	5.84e-04	1.33 (1.13-1.57)	66.8	4.91e-02
rs672757	2	genotypic	1.41e-03	1.23 (1.08-1.39)	90.2	1.40e-03
rs7600624	4	genotypic	1.89e-03	1.20 (1.07-1.34)	72.9	1.14e-02
rs4649314	4	recessive	3.50e-03	1.22 (1.07-1.39)	70.2	1.80e-02
rs567564	3	recessive	4.02e-03	1.37 (1.11-1.70)	81.2	4.95e-03
rs1034116	4	genotypic	4.24e-03	1.19 (1.06-1.34)	73.7	9.75e-03
rs10429965	3	recessive	4.45e-03	1.89 (1.22-2.92)	89.8	5.58e-05
rs1571583	4	recessive	5.66e-03	1.40 (1.10-1.78)	70.5	1.72e-02
rs6972789	3	recessive	6.95e-03	1.24 (1.06-1.46)	91.1	1.30e-05
rs1438620	2	dominant	7.04e-03	1.23 (1.06-1.43)	76.8	1.33e-02
rs437171	2	recessive	1.16e-02	1.46 (1.09-1.97)	76.6	1.39e-02
rs1822917	4	recessive	1.20e-02	1.39 (1.07-1.79)	59.2	6.13e-02
rs4715476	3	genotypic	1.37e-02	1.16 (1.03-1.31)	81.5	4.46e-03
rs712082	3	recessive	1.48e-02	1.43 (1.07-1.91)	86.6	5.75e-04
rs1350308	3	genotypic	1.61e-02	1.14 (1.02-1.27)	79.3	7.90e-03
rs10510044	2	recessive	1.66e-02	1.42 (1.07-1.89)	81.5	4.54e-03
rs7007146	3	recessive	3.00e-02	1.20 (1.02-1.41)	84.0	1.94e-03
rs5997921	2	recessive	3.14e-02	1.73 (1.05-2.86)	79.7	2.64e-02
rs9315425	3	recessive	3.25e-02	1.63 (1.04-2.55)	73.8	2.21e-02
rs7866165	3	recessive	4.30e-02	1.26 (1.01-1.57)	83.6	2.28e-03
rs11788150	3	recessive	4.46e-02	1.21 (1.00-1.46)	85.6	9.87e-04
rs6518956	4	genotypic	5.59e-02	0.92 (0.84-1.00)	0.0	6.00e-01
rs1824476	3	genotypic	6.93e-02	0.91 (0.81-1.01)	86.0	7.77e-04
rs3752261	3	recessive	7.24e-02	1.38 (0.97-1.97)	76.0	1.55e-02
rs10842099	2	genotypic	8.80e-02	0.88 (0.75-1.02)	90.6	1.13e-03
rs2514841	3	recessive	9.15e-02	0.86 (0.72-1.02)	80.8	5.50e-03
rs9514816	3	genotypic	1.10e-01	1.12 (0.97-1.29)	89.8	5.76e-05
rs4978394	3	genotypic	1.47e-01	1.10 (0.97-1.26)	85.1	1.21e-03
rs745888	3	recessive	1.69e-01	1.18 (0.93-1.50)	80.5	5.99e-03
rs4776494	2	genotypic	1.80e-01	1.12 (0.95-1.34)	88.2	3.60e-03
rs9533457	3	recessive	1.92e-01	1.18 (0.92-1.50)	86.7	5.58e-04
rs1924597	4	genotypic	2.48e-01	0.92 (0.81-1.06)	87.1	3.57e-05
rs25689	3	genotypic	3.82e-01	1.05 (0.94-1.18)	82.4	3.40e-03
rs2589183	4	genotypic	4.19e-01	0.96 (0.86-1.07)	83.4	4.35e-04
rs3784780	4	genotypic	9.11e-01	1.01 (0.90-1.13)	0.0	9.07e-01

Meta-analysis including stage in multivariate analysis, for patients of all stages.

Table 10-5 Meta-analysis for stage 2 and 3 CRC, adjusted for stage

SNP	Cohorts	Model	p-value	ES (95% CI)	I^2	p_Q
rs7556894	4	recessive	8.96E-07	1.80 (1.42-2.27)	56.9	7.31E-02
rs672757	2	allelic	1.09E-06	1.47 (1.26-1.72)	38.4	2.02E-01
rs764372	3	recessive	1.25E-05	2.30 (1.58-3.34)	66.4	5.09E-02
rs472660	3	recessive	1.79E-05	2.45 (1.63-3.70)	63.9	6.25E-02
rs10429965	3	recessive	8.52E-05	2.66 (1.63-4.33)	89.9	1.68E-03
rs7912136	2	recessive	1.22E-04	1.87 (1.36-2.56)	80.7	2.28E-02
rs1878632	3	recessive	3.78E-04	1.38 (1.16-1.65)	59.9	8.25E-02
rs7600624	4	allelic	6.79E-04	1.24 (1.10-1.41)	70.5	1.71E-02
rs437171	2	recessive	8.63E-04	2.13 (1.37-3.33)	72.8	5.53E-02
rs4649314	4	recessive	1.34E-03	1.29 (1.10-1.50)	64.4	3.78E-02
rs712082	3	recessive	1.34E-03	1.70 (1.23-2.34)	79.5	7.58E-03
rs6972789	3	recessive	1.59E-03	1.32 (1.11-1.58)	89.9	5.02E-05
rs1571583	4	recessive	1.76E-03	1.53 (1.17-2.00)	67.5	2.65E-02
rs7007146	3	recessive	2.22E-03	1.32 (1.10-1.57)	85.8	8.92E-04
rs5997921	2	recessive	2.59E-03	2.43 (1.36-4.34)	0.0	3.38E-01
rs10510044	2	recessive	3.77E-03	1.64 (1.17-2.29)	88.7	2.90E-03
rs1438620	2	dominant	4.10E-03	1.46 (1.13-1.89)	83.8	1.29E-02
rs567564	3	recessive	5.37E-03	1.40 (1.10-1.78)	81.8	4.12E-03
rs1822917	4	recessive	7.78E-03	1.47 (1.11-1.96)	54.6	8.53E-02
rs11788150	3	recessive	8.63E-03	1.31 (1.07-1.61)	83.2	2.59E-03
rs1350308	3	allelic	9.22E-03	1.20 (1.05-1.37)	75.8	1.62E-02
rs4978394	3	genotypic	1.15E-02	1.20 (1.04-1.39)	71.6	2.97E-02
rs9315425	3	recessive	1.19E-02	1.83 (1.14-2.94)	63.8	6.29E-02
rs7866165	3	recessive	1.35E-02	1.36 (1.07-1.74)	79.9	6.91E-03
rs1924597	4	allelic	2.13E-02	0.83 (0.71-0.97)	82.1	7.93E-04
rs10842099	2	allelic	2.64E-02	0.81 (0.67-0.98)	88.4	3.31E-03
rs4715476	3	genotypic	2.75E-02	1.16 (1.02-1.33)	82.8	2.95E-03
rs1034116	4	allelic	3.00E-02	1.16 (1.01-1.33)	76.3	5.40E-03
rs9514816	3	allelic	4.30E-02	1.17 (1.00-1.36)	88.6	1.51E-04
rs3752261	3	recessive	5.03E-02	1.47 (1.00-2.16)	75.4	1.72E-02
rs9533457	3	recessive	5.71E-02	1.28 (0.99-1.66)	82.2	3.68E-03
rs4776494	2	genotypic	5.89E-02	1.22 (0.99-1.50)	85.1	9.70E-03
rs25689	3	allelic	1.14E-01	1.11 (0.98-1.25)	75.4	1.71E-02
rs745888	3	recessive	1.57E-01	1.21 (0.93-1.59)	83.6	2.21E-03
rs2514841	3	recessive	1.69E-01	0.87 (0.72-1.06)	80.7	5.64E-03
rs1824476	3	allelic	2.71E-01	0.94 (0.83-1.05)	84.3	1.69E-03
rs6518956	3	genotypic	3.07E-01	0.94 (0.85-1.05)	0.0	6.28E-01
rs2589183	4	allelic	3.14E-01	0.94 (0.83-1.06)	82.4	6.97E-04
rs3784780	4	genotypic	6.30E-01	0.97 (0.85-1.11)	0.0	7.91E-01

Meta-analysis including stage in multivariate analysis, for patients with stage 2 and 3 CRC.

Table 10-6 Meta-analysis for stage 2 and 3, adjusted for stage, using shrunk data

SNP	Cohorts	Model	p-value	ES (95% CI)	I^2	p_Q
rs7556894	4	recessive	7.19e-04	1.56 (1.21-2.03)	0.0	6.23e-01
rs672757	2	allelic	2.52e-02	1.22 (1.03-1.46)	0.0	3.46e-01
rs764372	3	recessive	6.02e-02	1.44 (0.98-2.09)	0.0	5.27e-01
rs4649314	4	recessive	7.24e-02	1.15 (0.99-1.33)	0.0	5.08e-01
rs472660	3	recessive	9.04e-02	1.53 (0.94-2.49)	0.0	5.39e-01
rs7600624	4	allelic	1.15e-01	1.09 (0.98-1.20)	21.1	2.84e-01
rs1878632	3	recessive	1.18e-01	1.12 (0.97-1.30)	0.0	5.84e-01
rs6972789	3	recessive	1.83e-01	1.14 (0.94-1.38)	75.2	1.76e-02
rs7007146	3	recessive	2.98e-01	1.10 (0.92-1.31)	70.9	3.20e-02
rs1438620	2	dominant	2.98e-01	1.08 (0.93-1.25)	0.0	9.25e-01
rs1571583	4	recessive	3.18e-01	1.16 (0.87-1.56)	0.0	7.93e-01
rs4978394	3	genotypic	3.74e-01	1.05 (0.94-1.17)	0.0	9.67e-01
rs1924597	4	allelic	3.88e-01	0.93 (0.79-1.10)	0.0	4.40e-01
rs6518956	4	genotypic	4.31e-01	0.98 (0.92-1.04)	0.0	5.18e-01
rs7866165	3	recessive	5.25e-01	1.08 (0.85-1.38)	0.0	9.63e-01
rs7912136	2	recessive	5.27e-01	1.15 (0.75-1.77)	0.0	7.10e-01
rs1822917	4	recessive	5.27e-01	1.08 (0.85-1.37)	0.0	7.26e-01
rs25689	3	allelic	6.14e-01	1.02 (0.94-1.11)	0.0	8.15e-01
rs2589183	4	allelic	6.15e-01	1.03 (0.91-1.17)	0.0	8.24e-01
rs11788150	3	recessive	6.21e-01	1.06 (0.84-1.34)	0.0	5.02e-01
rs712082	3	recessive	6.29e-01	1.10 (0.75-1.63)	0.0	7.71e-01
rs745888	3	recessive	6.41e-01	0.94 (0.73-1.21)	18.7	2.92e-01
rs10429965	3	recessive	6.46e-01	1.17 (0.60-2.26)	0.0	3.80e-01
rs9514816	3	allelic	6.48e-01	0.96 (0.81-1.14)	0.0	6.07e-01
rs437171	2	recessive	6.56e-01	1.08 (0.78-1.49)	0.0	9.77e-01
rs567564	3	recessive	6.75e-01	1.06 (0.80-1.41)	0.0	6.29e-01
rs4776494	2	genotypic	6.90e-01	1.02 (0.92-1.13)	14.4	2.80e-01
rs10510044	2	recessive	7.13e-01	0.92 (0.61-1.41)	0.0	4.10e-01
rs5997921	2	recessive	7.34e-01	1.02 (0.90-1.17)	0.0	5.10e-01
rs1350308	3	allelic	7.37e-01	1.03 (0.88-1.20)	0.0	8.52e-01
rs9315425	3	recessive	7.61e-01	1.03 (0.87-1.20)	0.0	4.16e-01
rs1824476	3	allelic	7.71e-01	0.99 (0.91-1.07)	52.2	1.23e-01
rs1034116	4	allelic	7.71e-01	1.02 (0.89-1.17)	0.0	5.19e-01
rs4715476	3	genotypic	7.75e-01	1.02 (0.89-1.17)	0.0	8.26e-01
rs2514841	3	recessive	7.79e-01	0.98 (0.87-1.11)	0.0	7.36e-01
rs3784780	4	genotypic	8.04e-01	0.99 (0.93-1.06)	0.0	7.76e-01
rs3752261	3	recessive	9.17e-01	1.01 (0.82-1.24)	0.0	7.10e-01
rs9533457	3	recessive	9.20e-01	0.99 (0.79-1.23)	0.0	7.81e-01
rs10842099	2	allelic	9.20e-01	1.01 (0.83-1.23)	0.0	4.39e-01

Meta-analysis including stage in multivariate analysis, for patients with stage 2 and 3 CRC, the data for VICTOR has been regressed to the mean.

Table 10-7 Predictive analysis of top 40 SNPs

SNP	Position	HR (95% CI)	p-value	Gene
rs4715476	54.73	1.89 (1.61-2.17)	7.17e-06	Tubulointerstitial nephritis antigen (370kb)
rs6117279	0.65	1.72 (1.48-1.97)	1.04e-05	Sulfiredoxin 1 homolog (60kb)
rs1408811	12.57	2.22 (1.86-2.58)	1.39e-05	Tyrosinase-related protein 1(120kb)
rs917205	87.56	2.32 (1.94-2.71)	1.52e-05	ADAM metallopeptidase domain 22
rs1516708	92.72	0.57 (0.31-0.83)	1.87e-05	*FAM190A* transcript variant 1, mRNA
rs968757	120.98	1.69 (1.45-1.93)	2.02e-05	Phosphodiesterase 5A isoform (209kb)
rs7924634	132.93	1.91 (1.61-2.21)	2.27e-05	Opioid binding protein (25kb)
rs5743291	49.31	2.17 (1.81-2.52)	2.29e-05	Nucleotide-binding oligomerization
rs7787525	99.33	2.42 (2.01-2.83)	2.33e-05	Tripartite motif protein TRIM4 isoform
rs510902	238.99	1.68 (1.44-1.92)	2.58e-05	Ankyrin repeat/SOCS box-containing (8kb)
rs746017	130.96	0.51 (0.19-0.82)	2.61e-05	Podocalyxin-like isoform 1 (64kb)
rs3124238	89.44	1.66 (1.42-1.89)	3.05e-05	Death-associated protein kinase 1
rs10281994	125.27	1.90 (1.60-2.21)	3.19e-05	Glutamate receptor, isoform c (590kb)
rs2002059	101.12	2.32 (1.92-2.71)	3.53e-05	Cyclin M1
rs2144749	168.70	1.67 (1.43-1.91)	3.61e-05	SPARC related modular calcium binding
rs3784929	74.23	1.82 (1.54-2.11)	3.77e-05	Lysyl-tRNA synthetase isoform 3
rs10102339	27.02	1.66 (1.42-1.91)	3.89e-05	Stathmin-like 4 (126kb)
rs13253769	100.10	0.59 (0.34-0.84)	4.08e-05	Vacuolar protein sorting 13B isoform 4
rs1041795	38.77	1.70 (1.44-1.95)	4.38e-05	ETS-related isoform 1
rs8053257	74.23	2.03 (1.69-2.37)	4.94e-05	Lysyl-tRNA synthetase isoform 2
rs883834	45.31	2.07 (1.72-2.43)	5.01e-05	Homo sapiens cDNA FLJ38231
rs9514816	107.58	1.89 (1.58-2.20)	5.78e-05	DNA ligase IV (75kb)
rs4923791	36.48	1.97 (1.64-2.31)	6.41e-05	Sprouty-related protein 1 (90kb)
rs10913063	173.94	1.98 (1.64-2.31)	6.46e-05	Tenascin R
rs237176	26.57	0.55 (0.26-0.84)	6.79e-05	Heparan sulfate D-glucosaminyl (510kb)
rs7591253	34.84	1.67 (1.42-1.93)	7.06e-05	Full-length cDNA clone CS0DI009YL06 (39kb)
rs2013746	7.33	1.79 (1.50-2.08)	7.35e-05	Ring finger protein 144 (227kb)
rs337532	100.49	2.21 (1.82-2.61)	7.52e-05	G protein-coupled receptor 52
rs1819887	50.65	1.67 (1.42-1.93)	7.90e-05	RAB27B, member RAS oncogene family
rs6491307	96.25	0.47 (0.09-0.85)	8.10e-05	Heparan sulfate 6-O-sulfotransferase 3
rs10934005	111.19	1.71 (1.44-1.97)	8.42e-05	Developmental pluripotency associated (645kb)
rs5764560	42.88	0.57 (0.29-0.85)	8.53e-05	Parvin, beta isoform a
rs10161354	87.87	2.25 (1.84-2.65)	8.73e-05	cDNA FLJ30500 (60kb)
rs12419726	59.50	1.80 (1.50-2.09)	9.17e-05	mRNA for hypothetical protein (0.5kb)
rs17596719	26.21	1.93 (1.60-2.26)	9.49e-05	Splice variant delE3-7 (HFE) (0.1kb)
rs1892431	90.14	1.81 (1.51-2.11)	9.61e-05	Leucine rich repeat containing 8 family
rs1730265	49.84	1.88 (1.56-2.20)	9.66e-05	Methyl-CpG binding domain protein 2 (90kb)
rs11709756	103.33	1.89 (1.57-2.21)	9.90e-05	Zona pellucida-like domain containing 1

Appendix B: Plots

Plots for top 10 Eigenvectors

Figure 10-1 Comparison of Eigenvectors 2 and 3

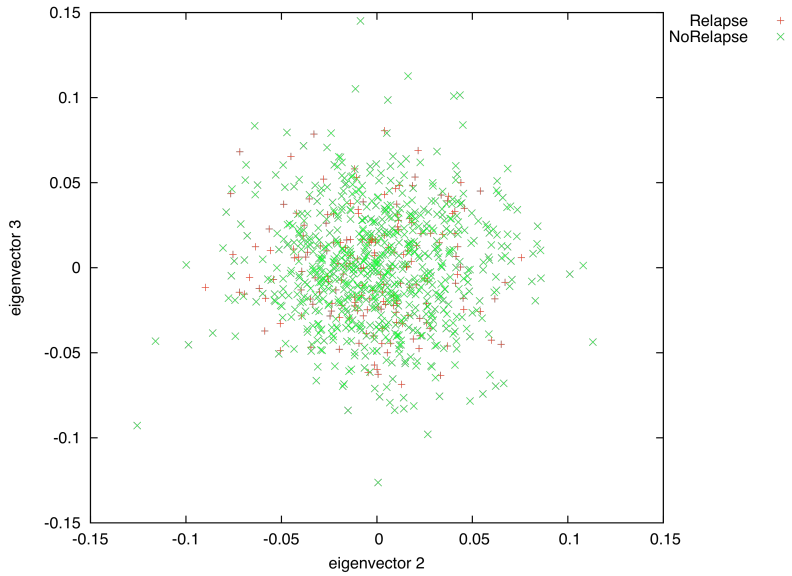


Figure 10-2 Comparison of Eigenvectors 3 and 4

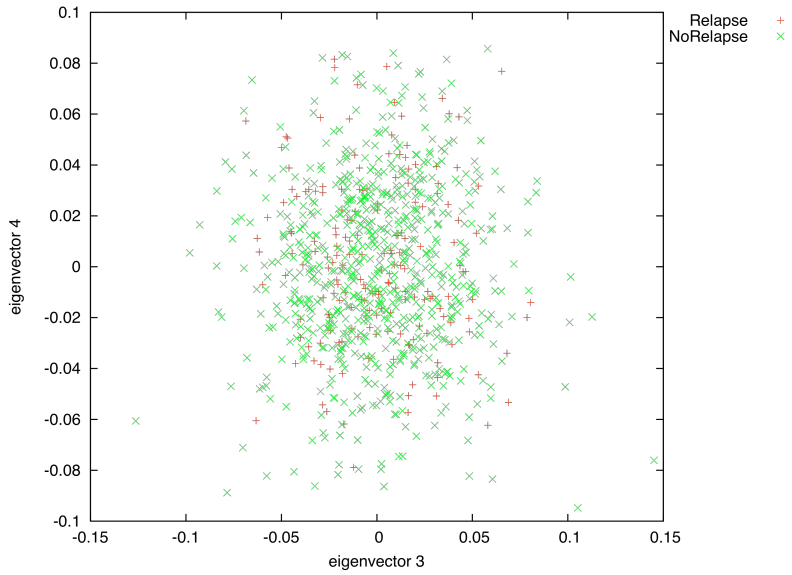


Figure 10-3 Comparison of Eigenvectors 4 and 5

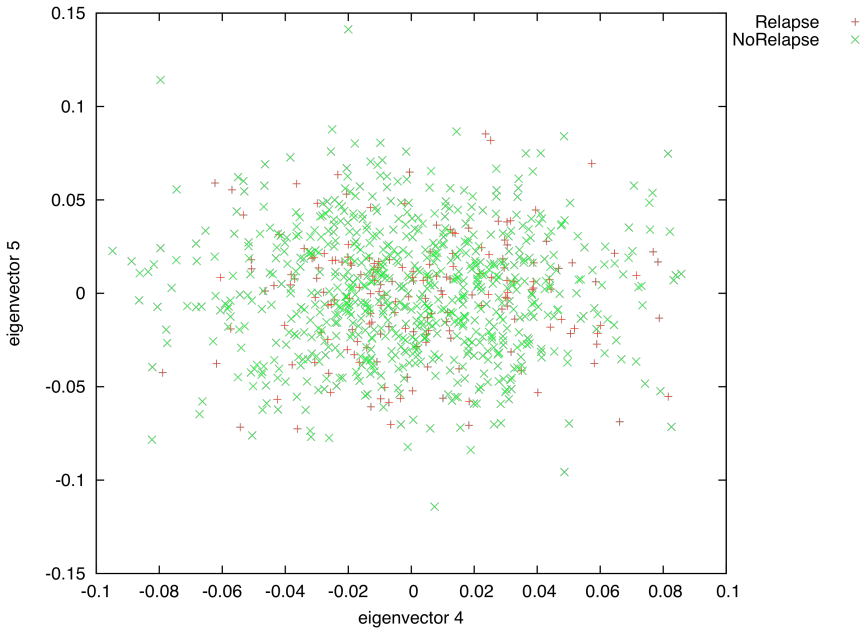


Figure 10-4 Comparison of Eigenvectors 5 and 6

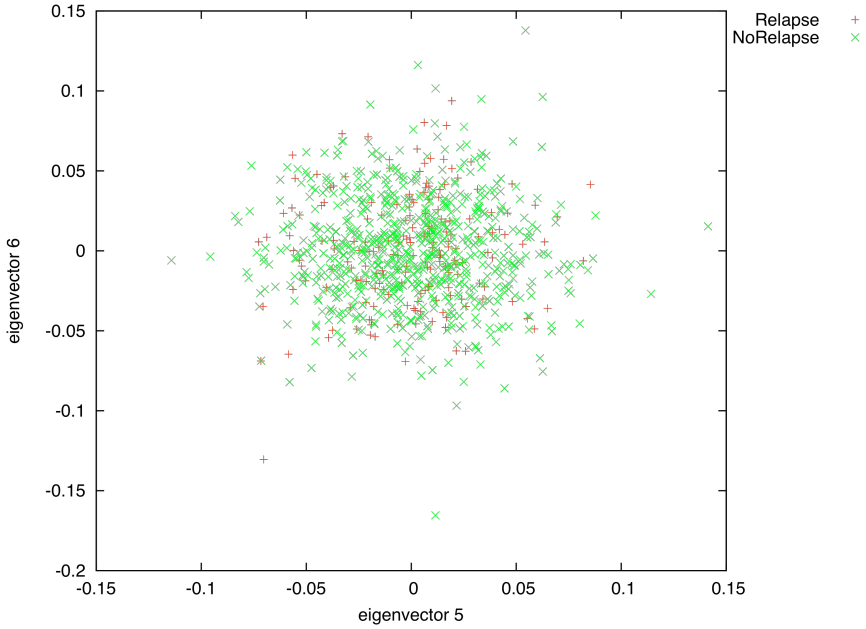


Figure 10-5 Comparison of Eigenvectors 6 and 7

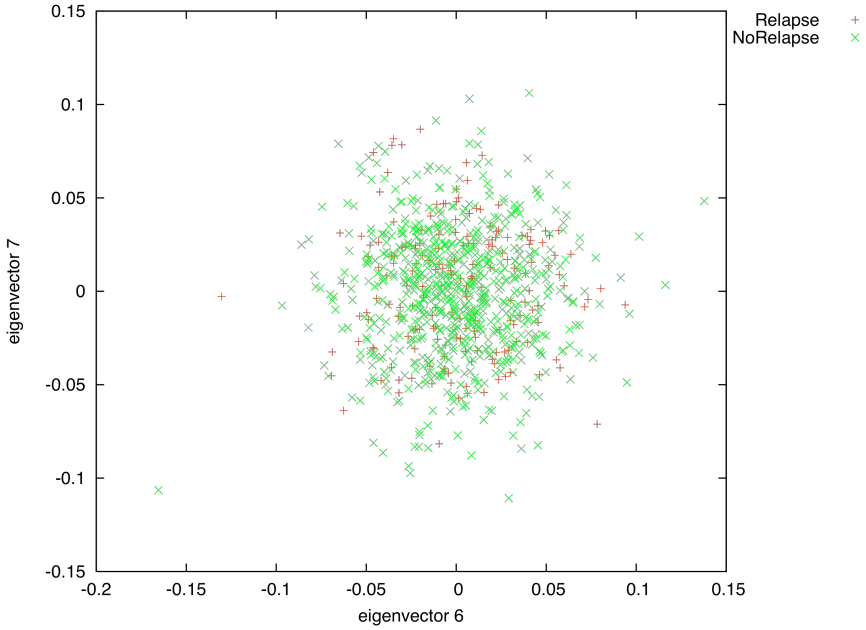


Figure 10-6 Comparison of Eigenvectors 7 and 8

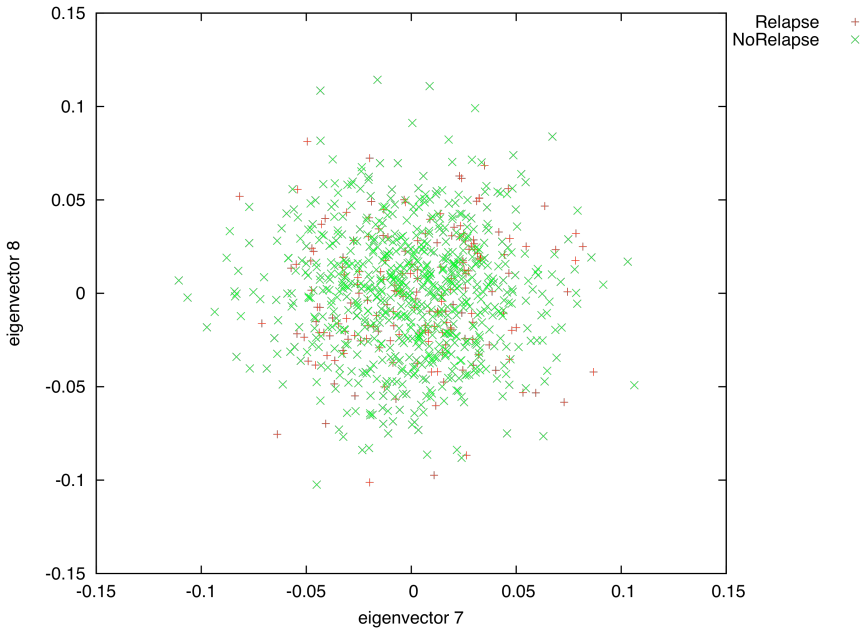


Figure 10-7 Comparison of Eigenvectors 8 and 9

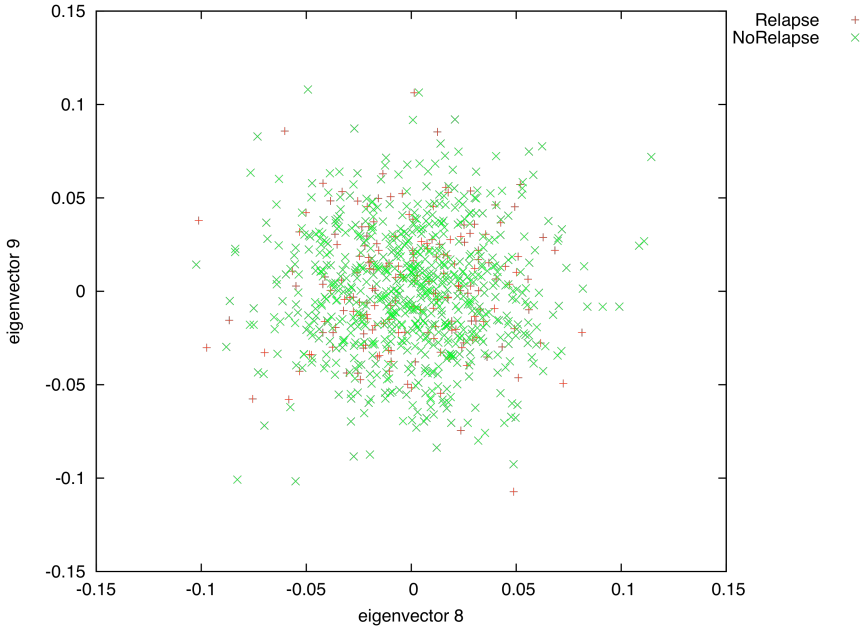
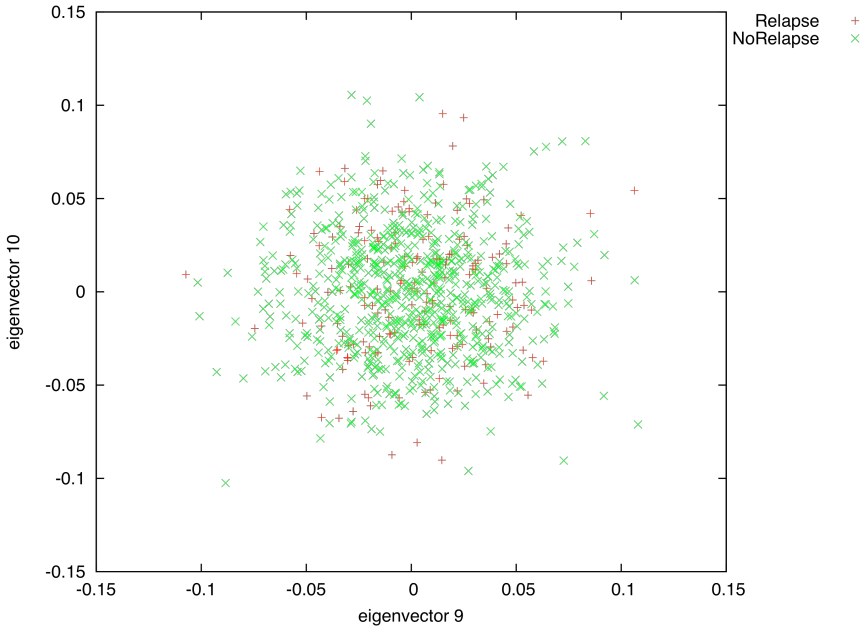


Figure 10-8 Comparison of Eigenvectors 9 and 10



Kaplan-Meier curves for top 40 SNPs

Figure 10-9 Kaplan-Meier curve for rs25689

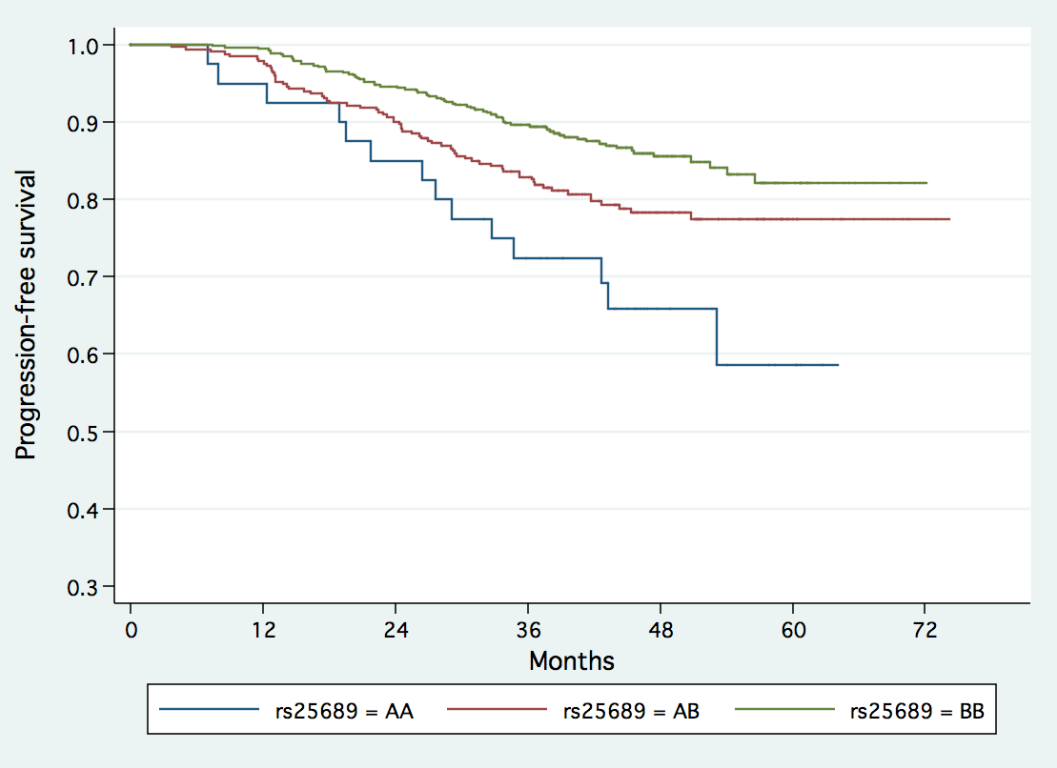


Figure 10-10 Kaplan-Meier curve for rs437171

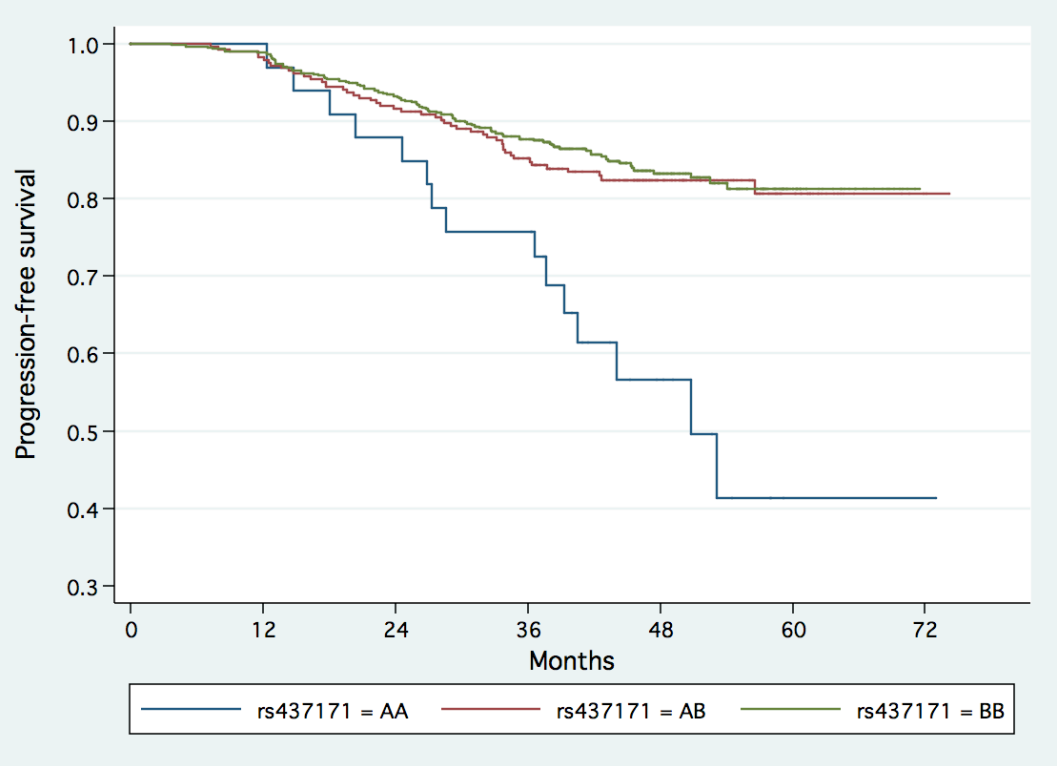


Figure 10-11 Kaplan-Meier curve for rs472660

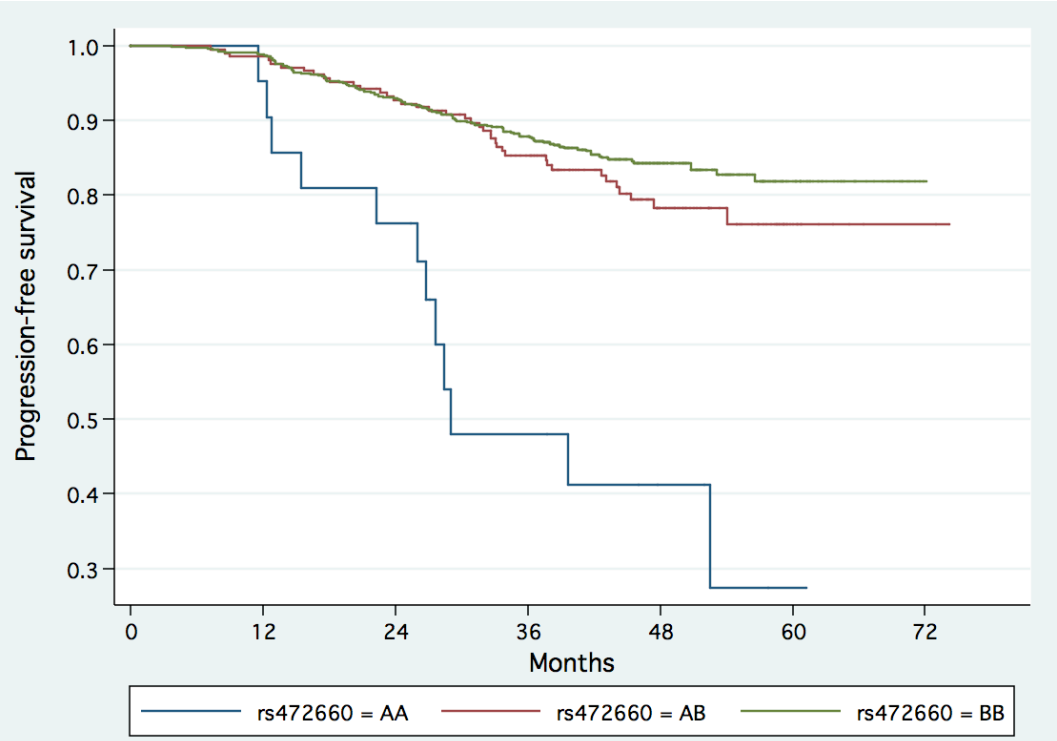


Figure 10-12 Kaplan-Meier curve for rs567564

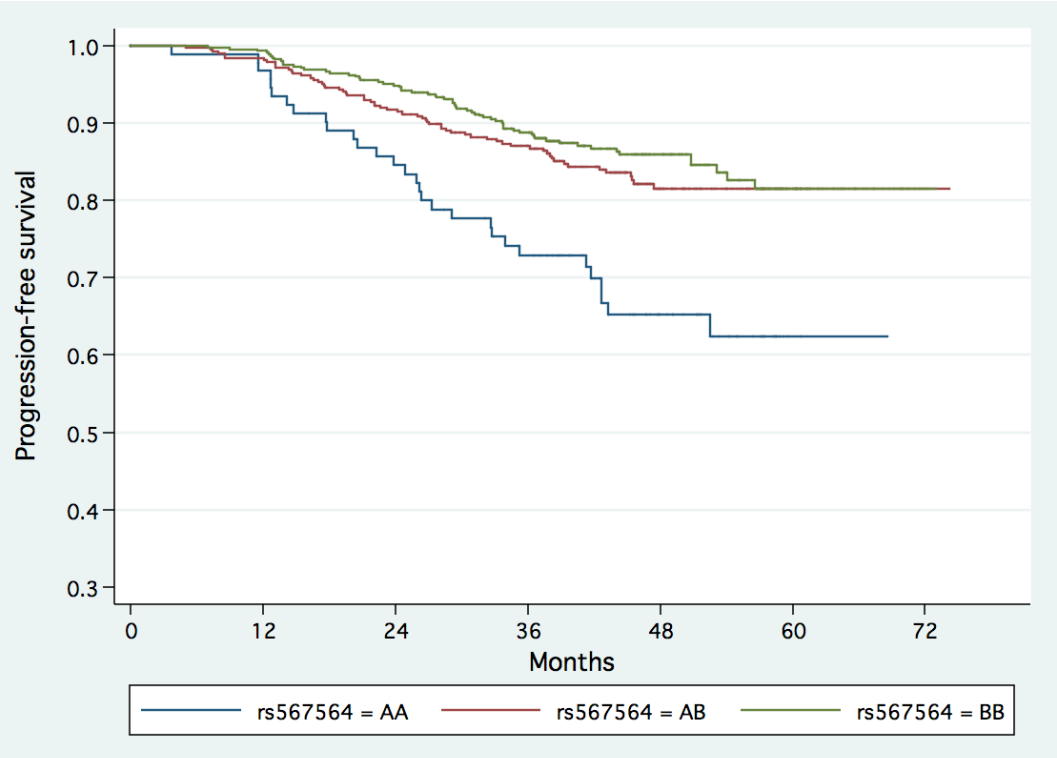


Figure 10-13 Kaplan-Meier curve for rs672757

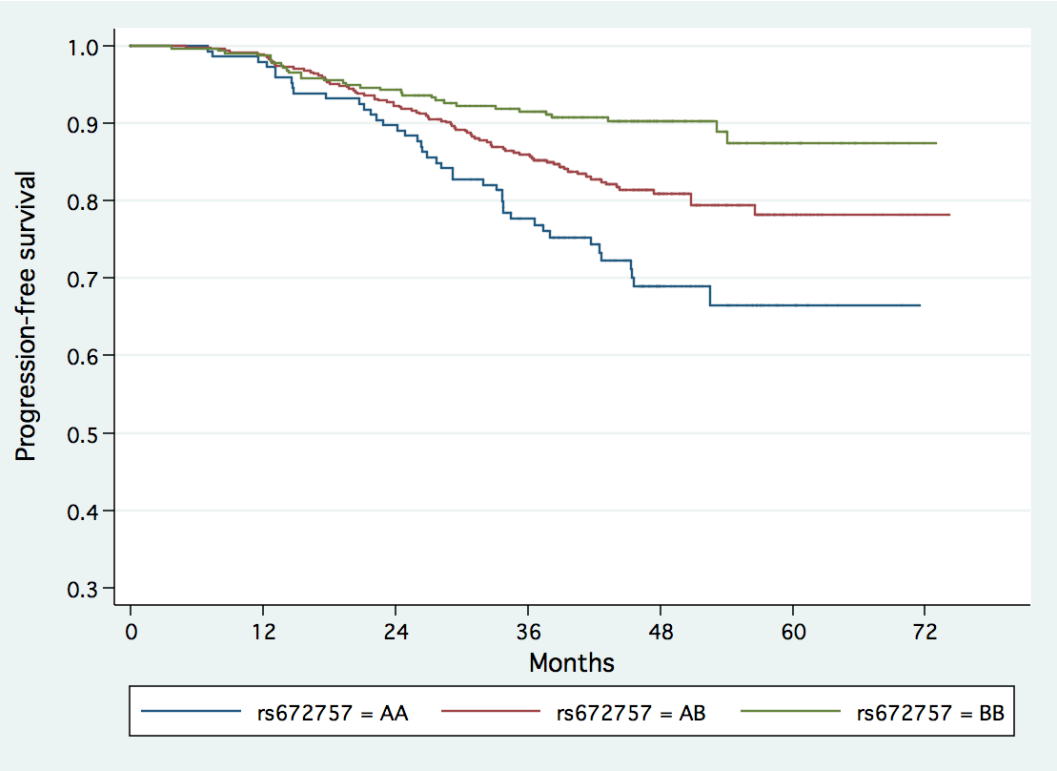


Figure 10-14 Kaplan-Meier curve for rs712082

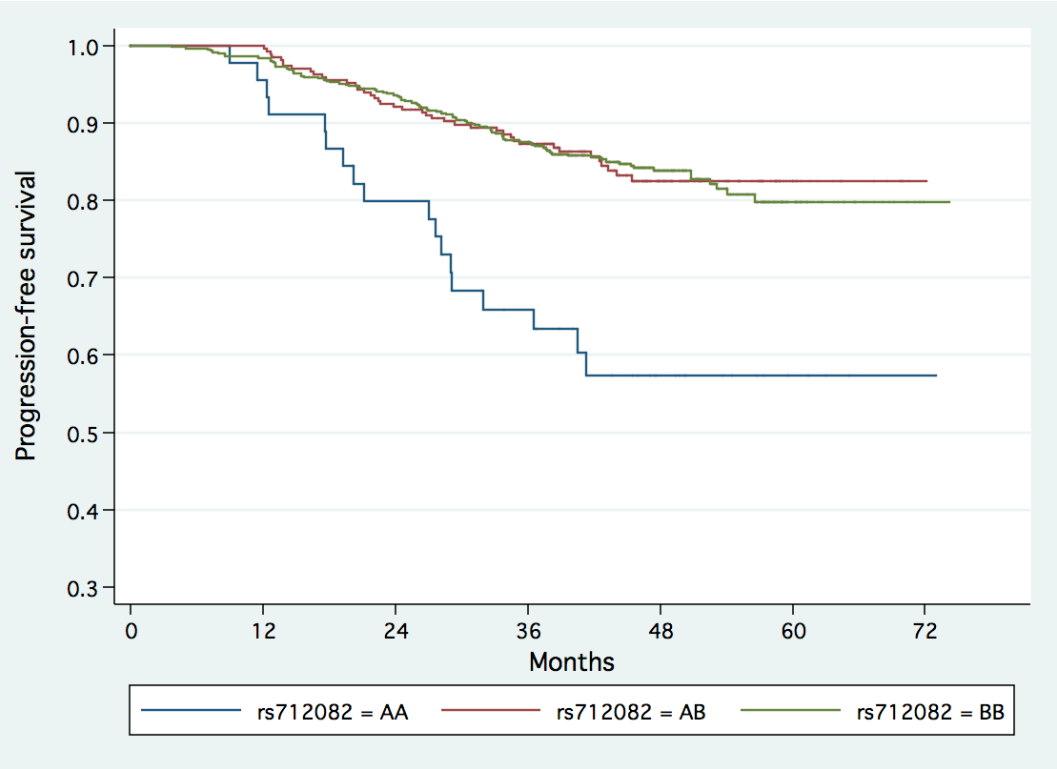


Figure 10-15 Kaplan-Meier curve for rs745888

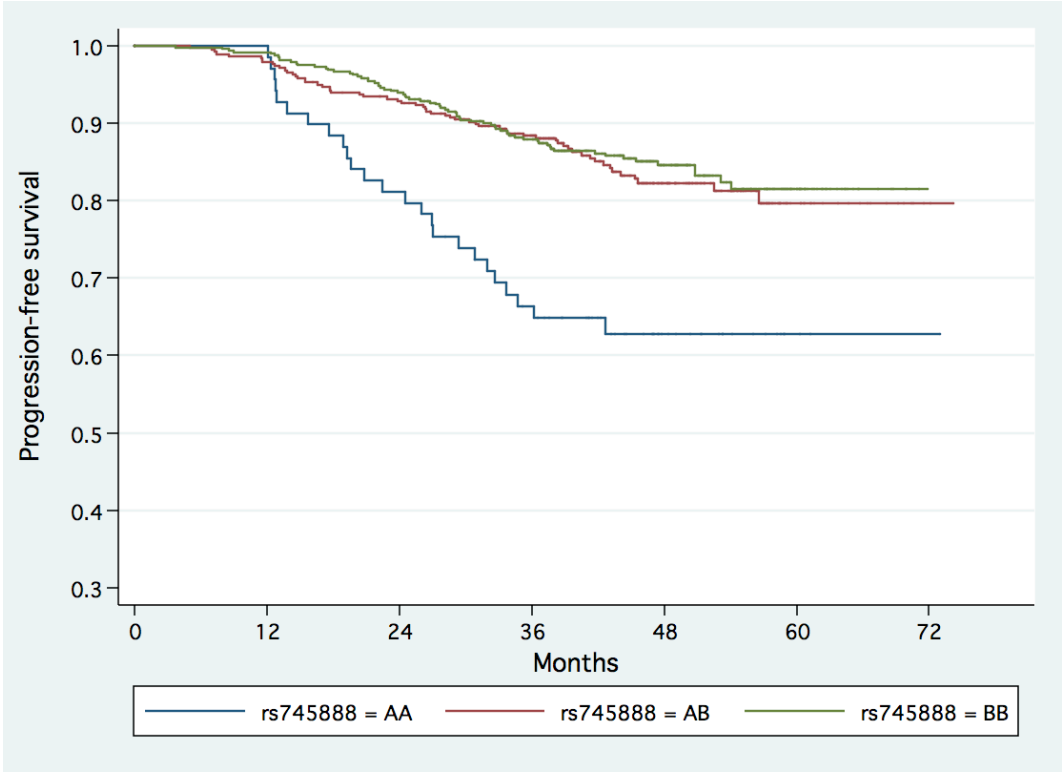


Figure 10-16 Kaplan-Meier curve for rs764372

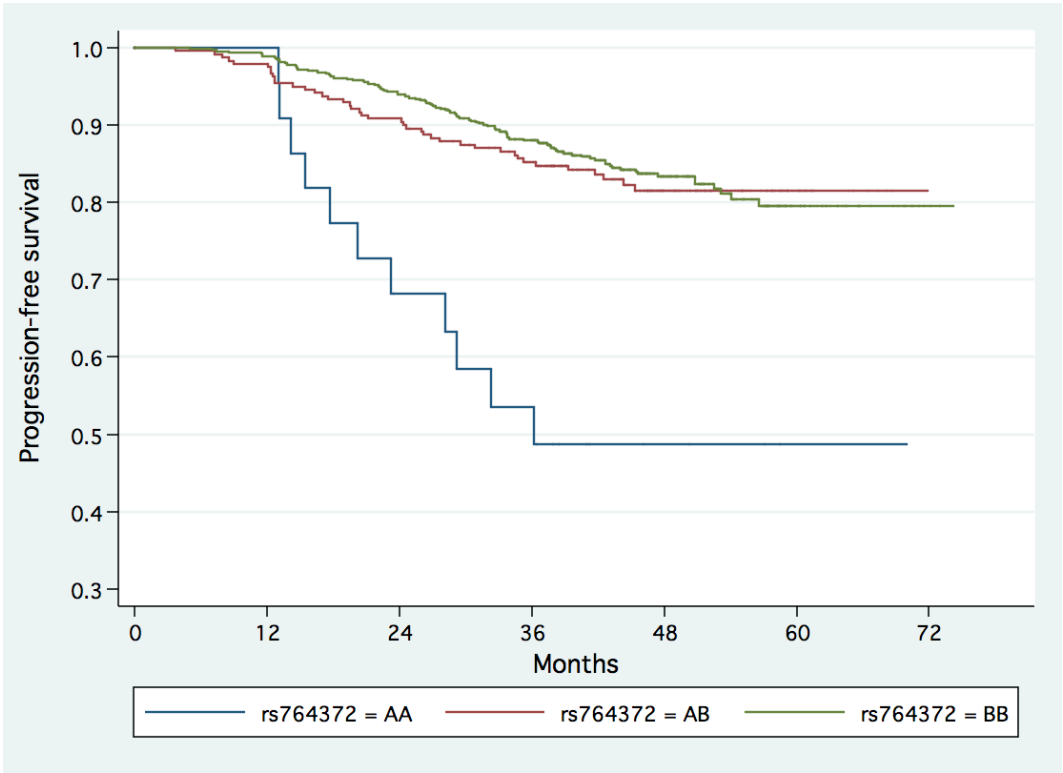


Figure 10-17 Kaplan-Meier curve for rs1034116

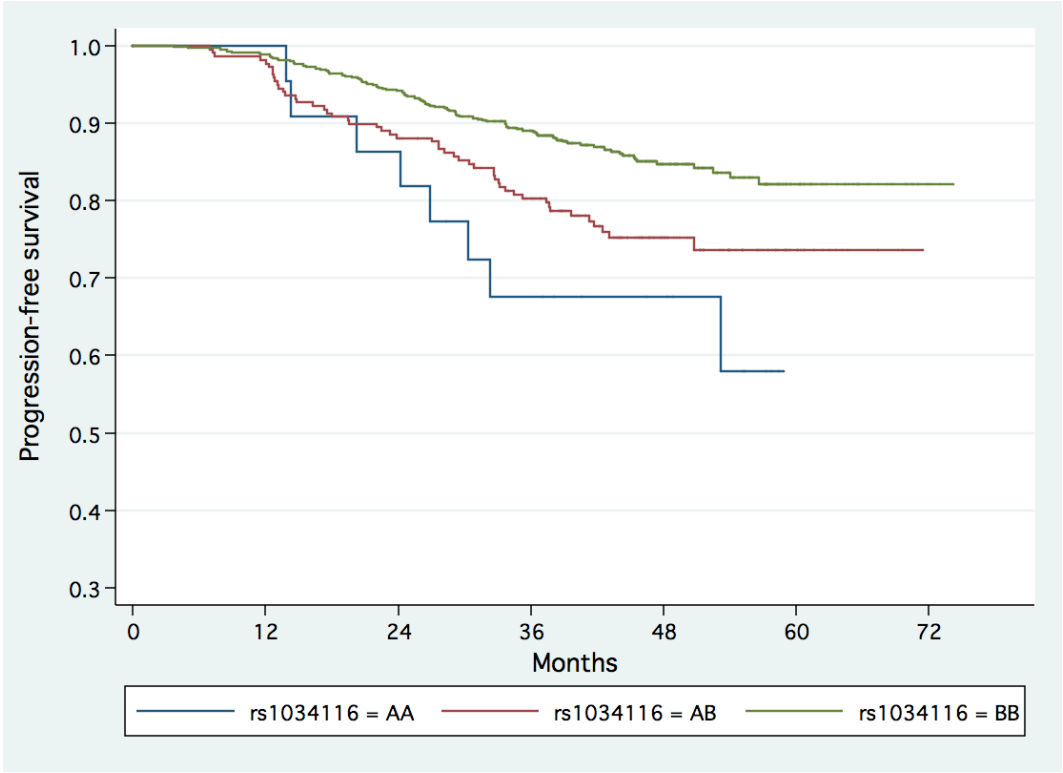


Figure 10-18 Kaplan-Meier curve for rs1350308

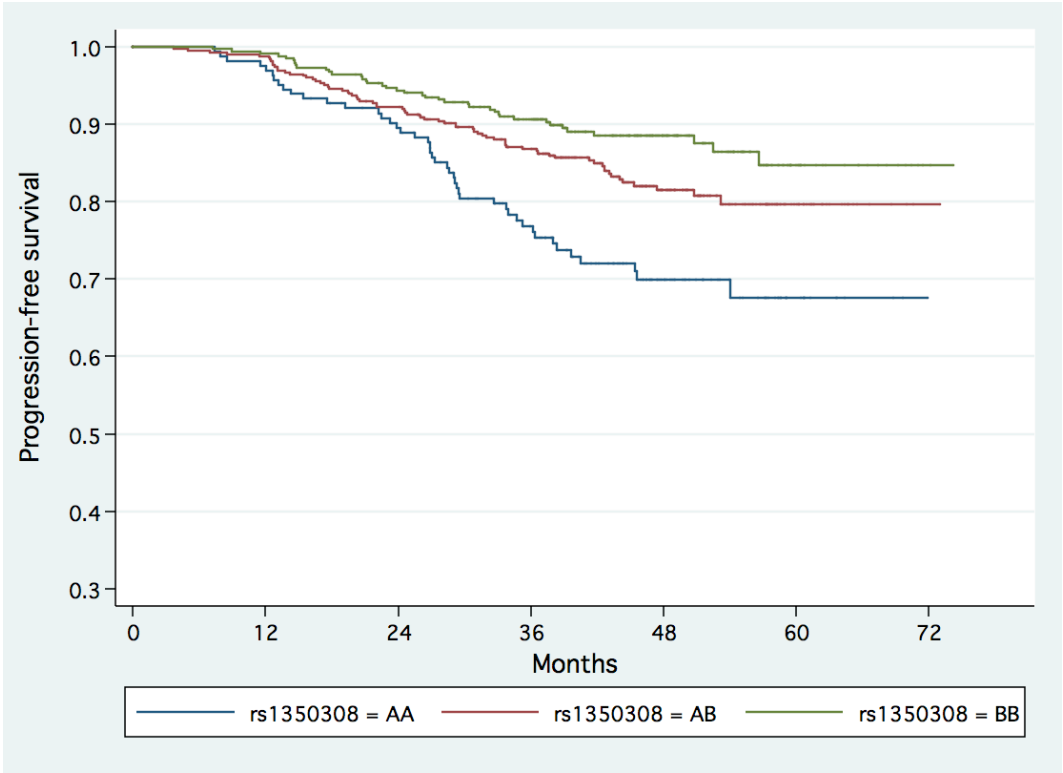


Figure 10-19 Kaplan-Meier curve for rs148620

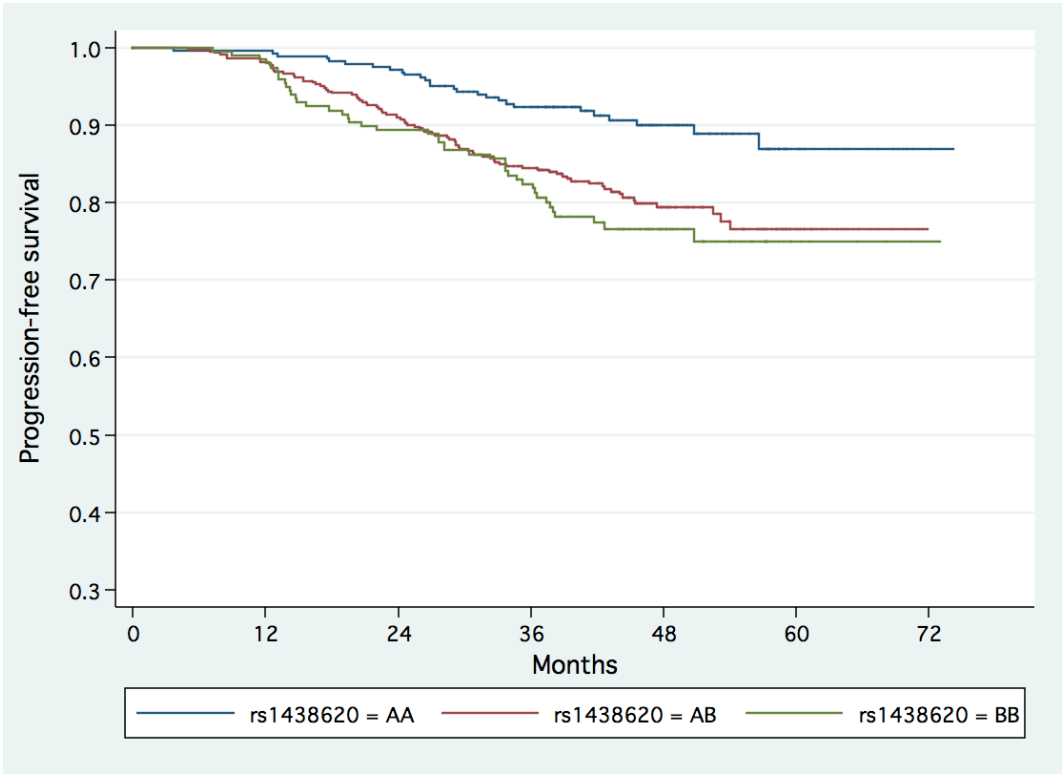


Figure 10-20 Kaplan-Meier curve for rs1526884

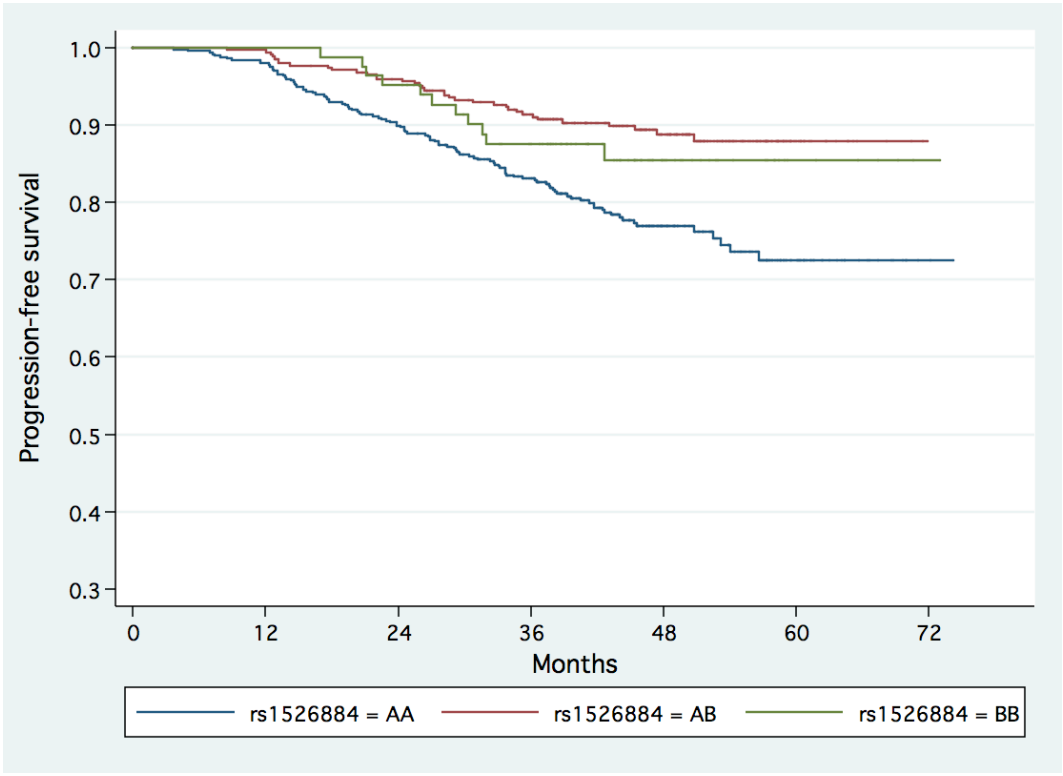


Figure 10-21 Kaplan-Meier curve for rs1571583

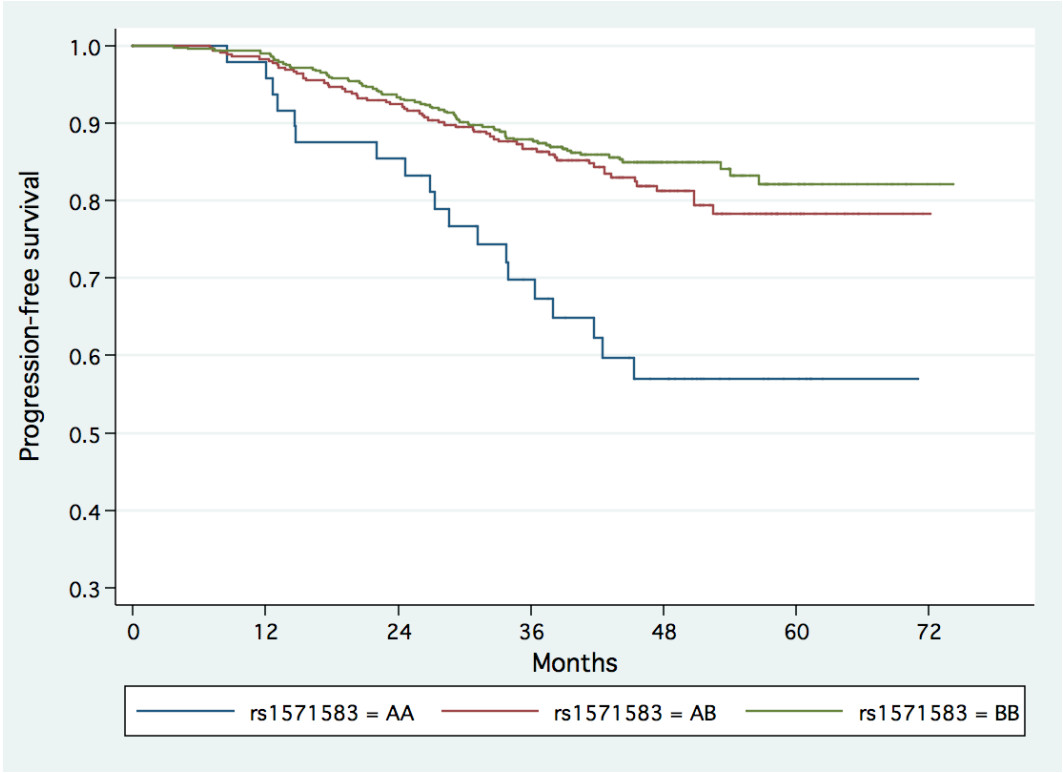


Figure 10-22 Kaplan-Meier curve for rs1822917

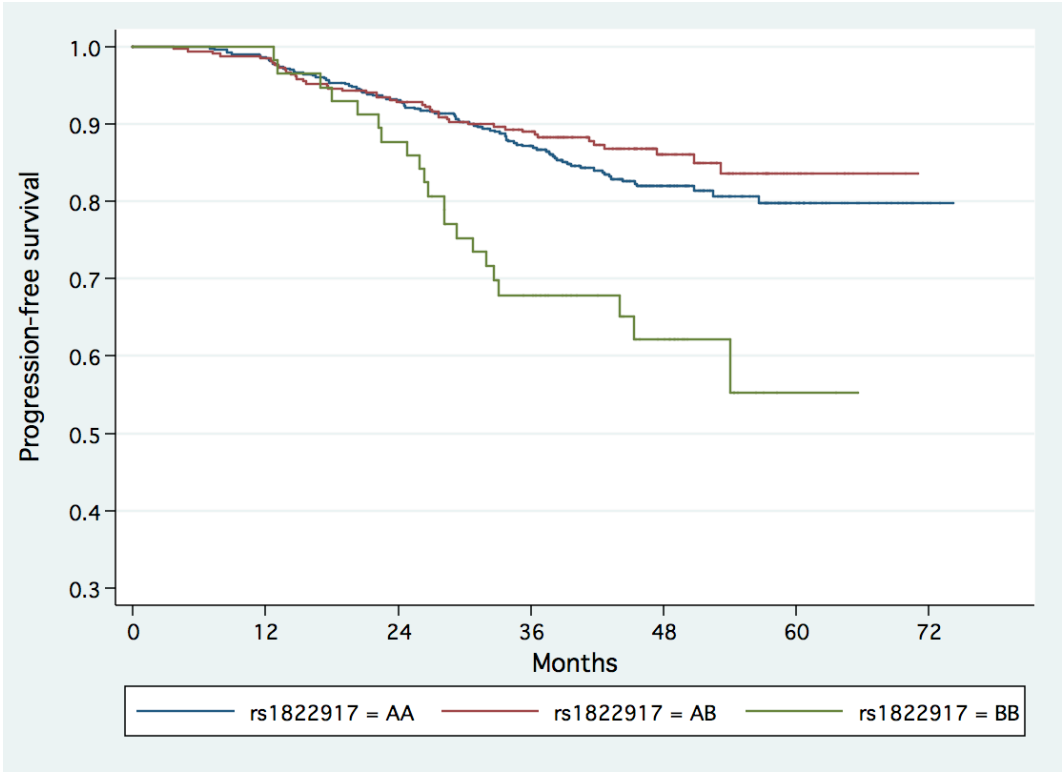


Figure 10-23 Kaplan-Meier curve for rs1824476

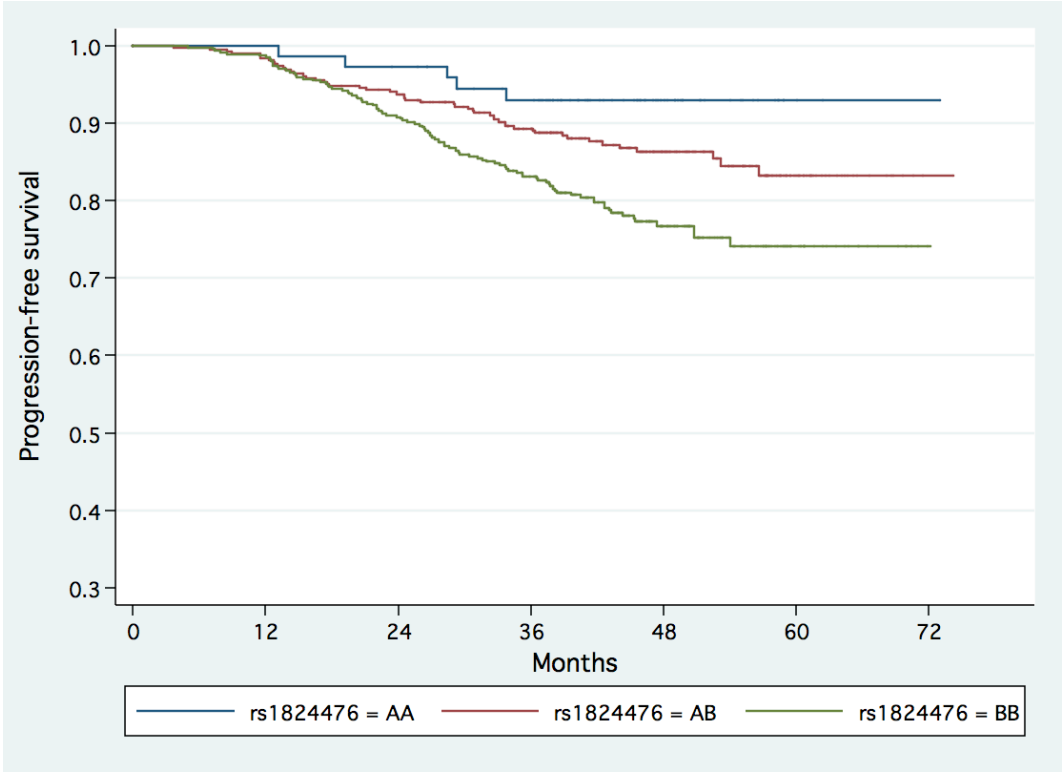


Figure 10-24 Kaplan-Meier curve for rs1878632

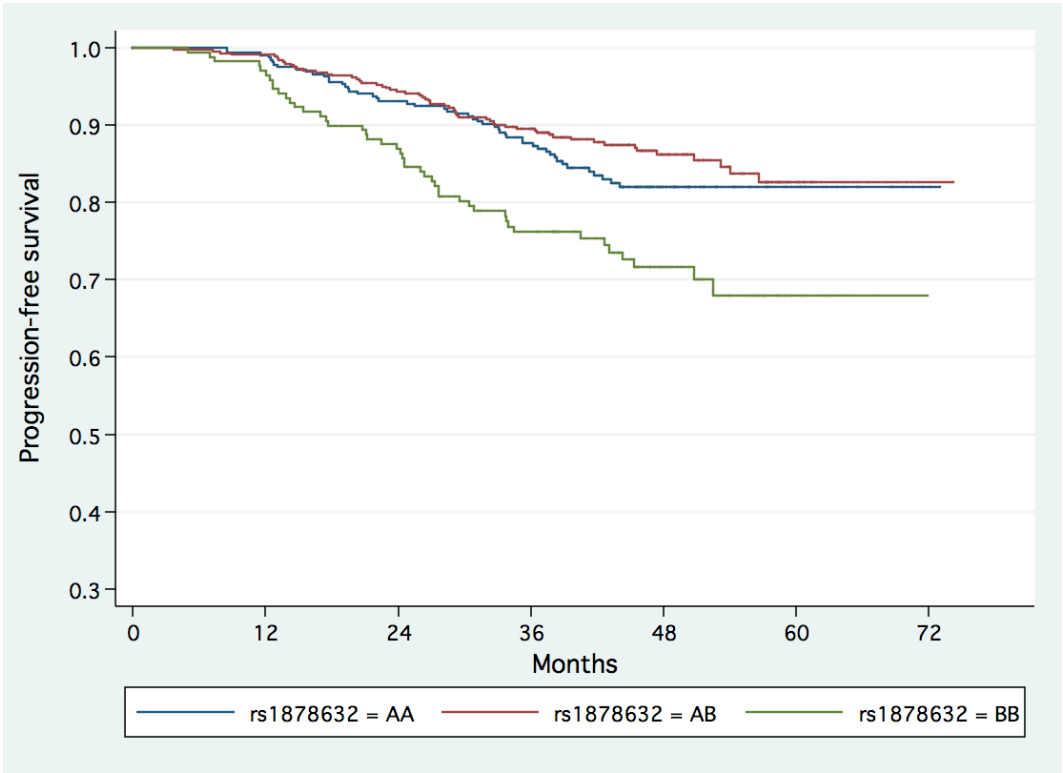


Figure 10-25 Kaplan-Meier curve for rs1924597

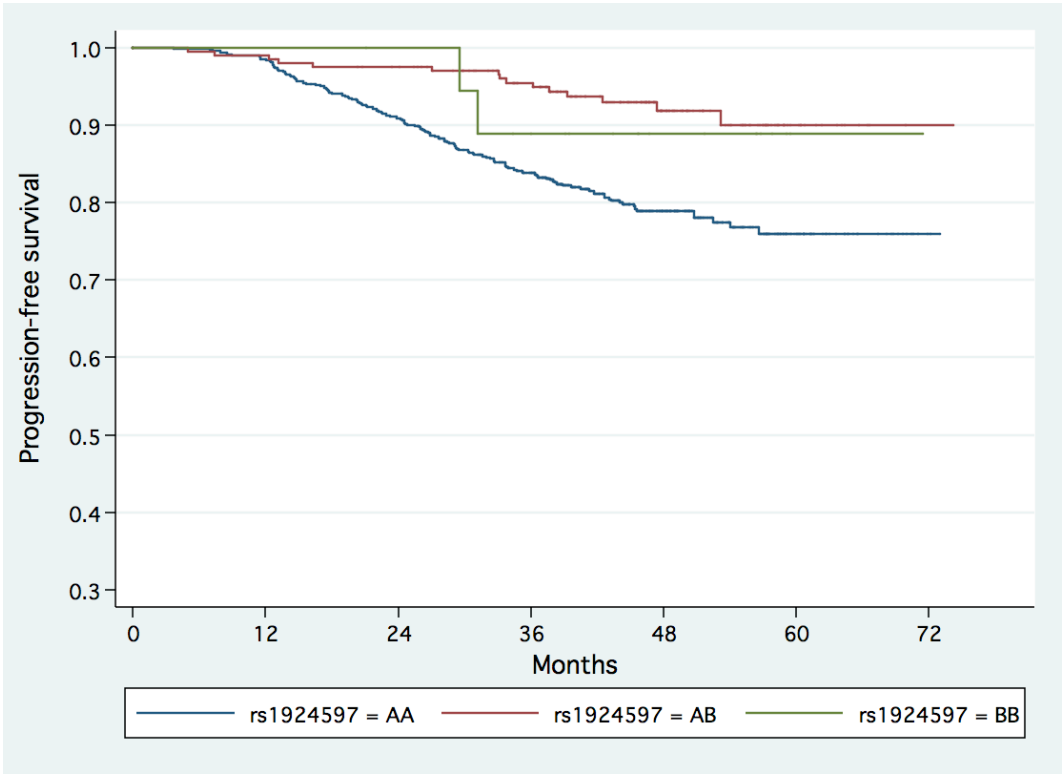


Figure 10-26 Kaplan-Meier curve for rs2514841

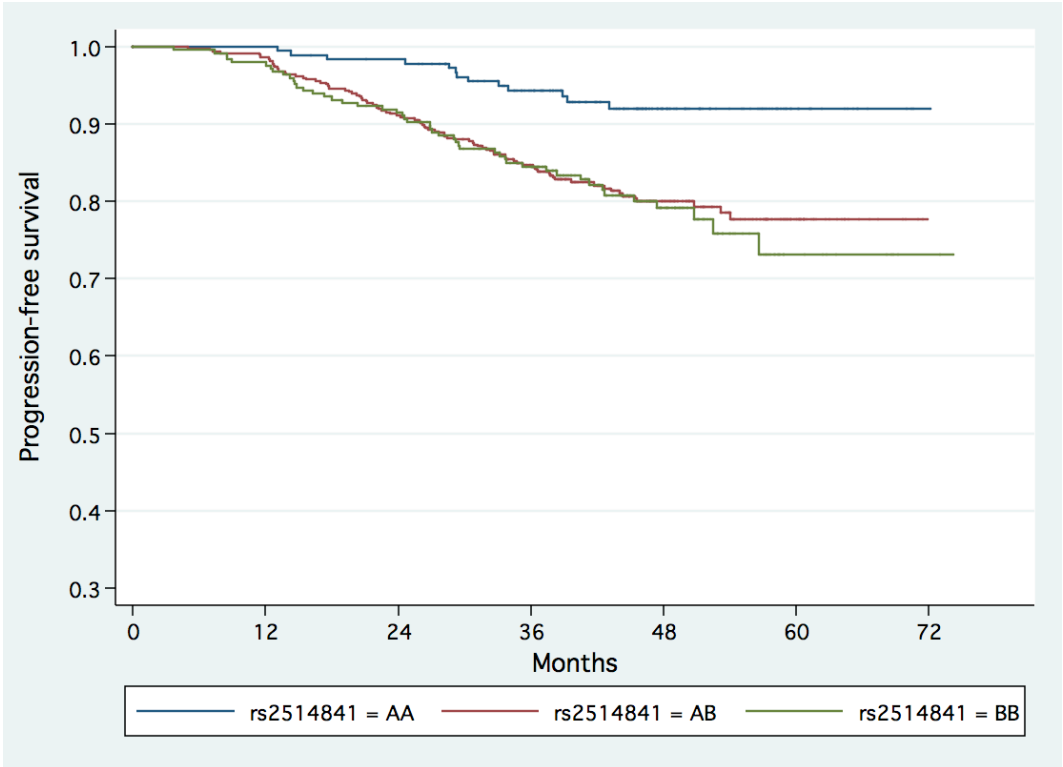


Figure 10-27 Kaplan-Meier curve for rs2589183

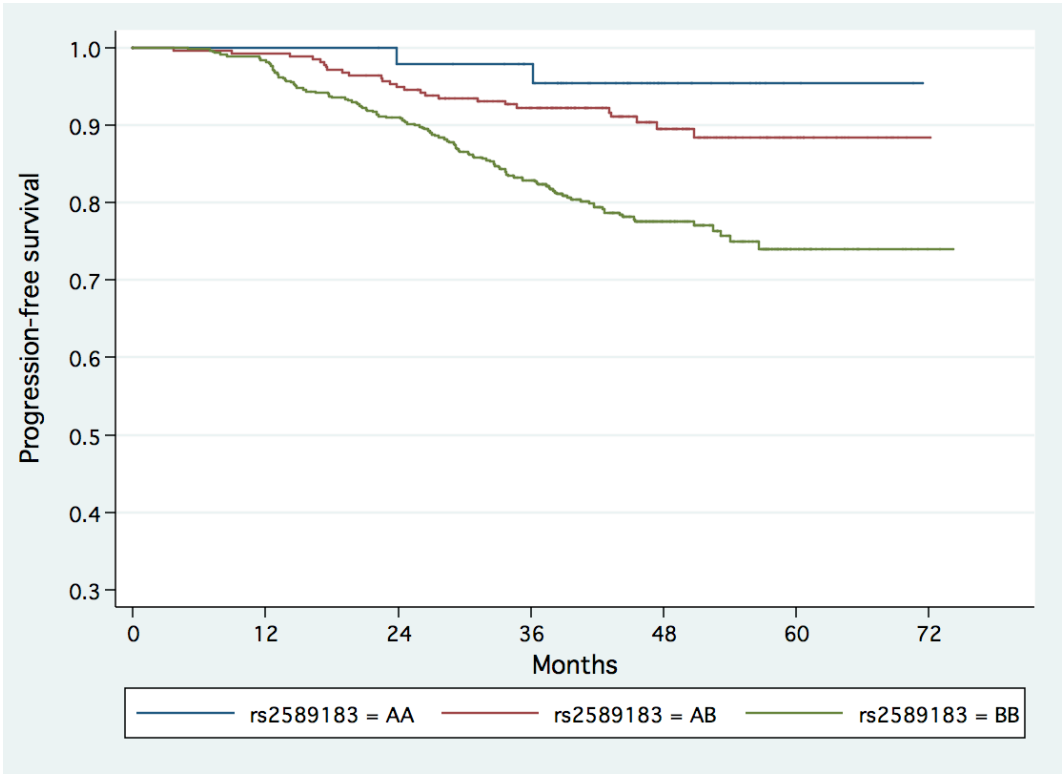


Figure 10-28 Kaplan-Meier curve for rs3752261

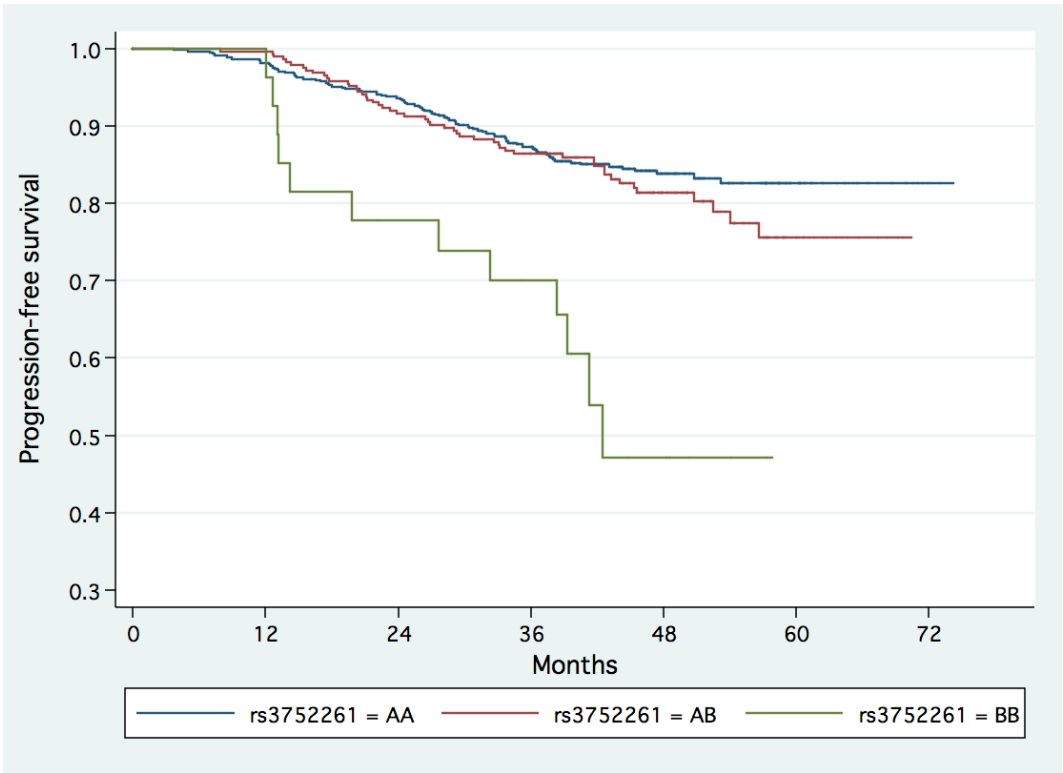


Figure 10-29 Kaplan-Meier curve for rs3784780

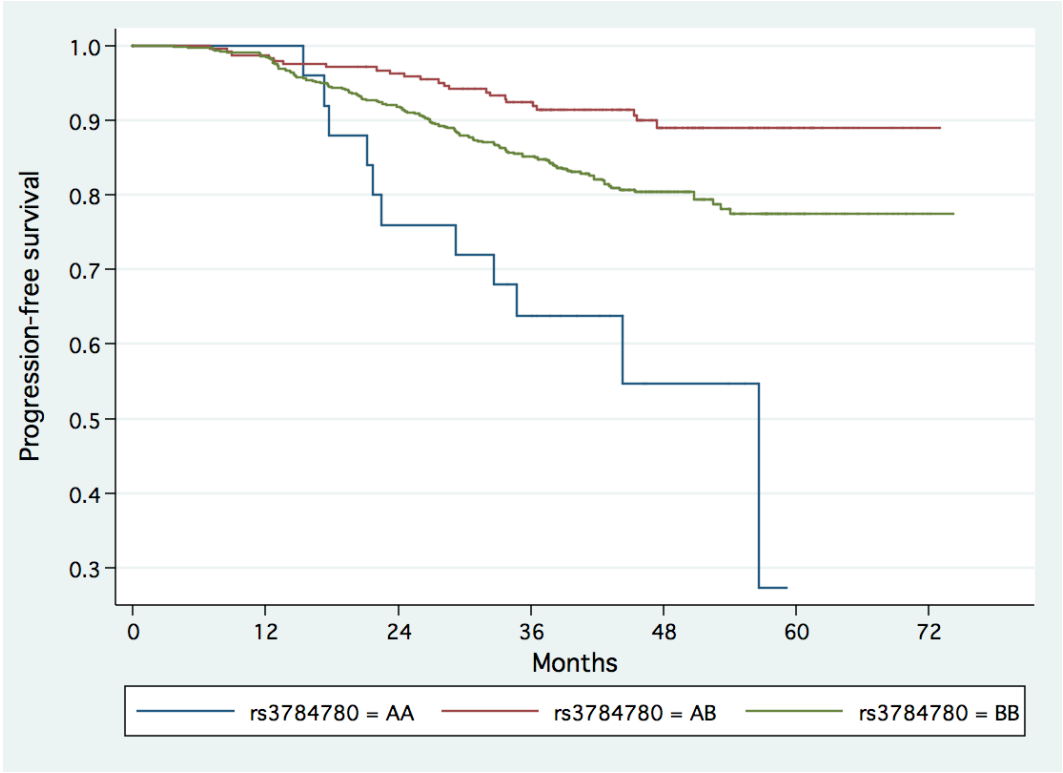


Figure 10-30 Kaplan-Meier curve for rs4649314

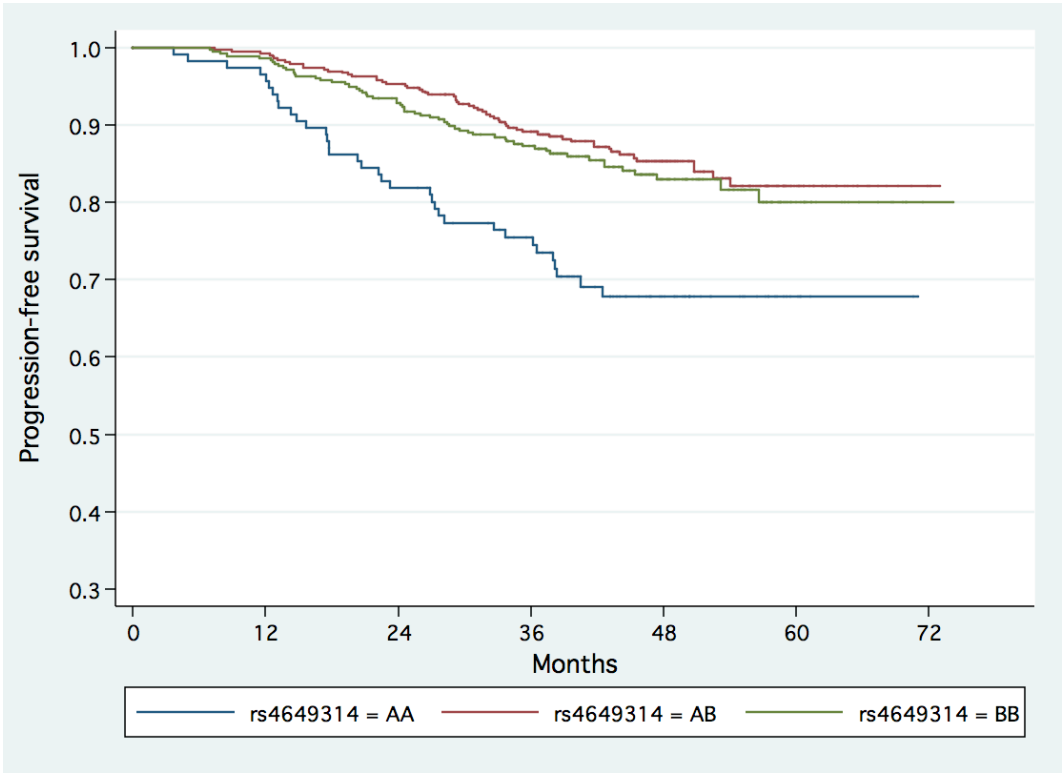


Figure 10-31 Kaplan-Meier curve for rs4715476

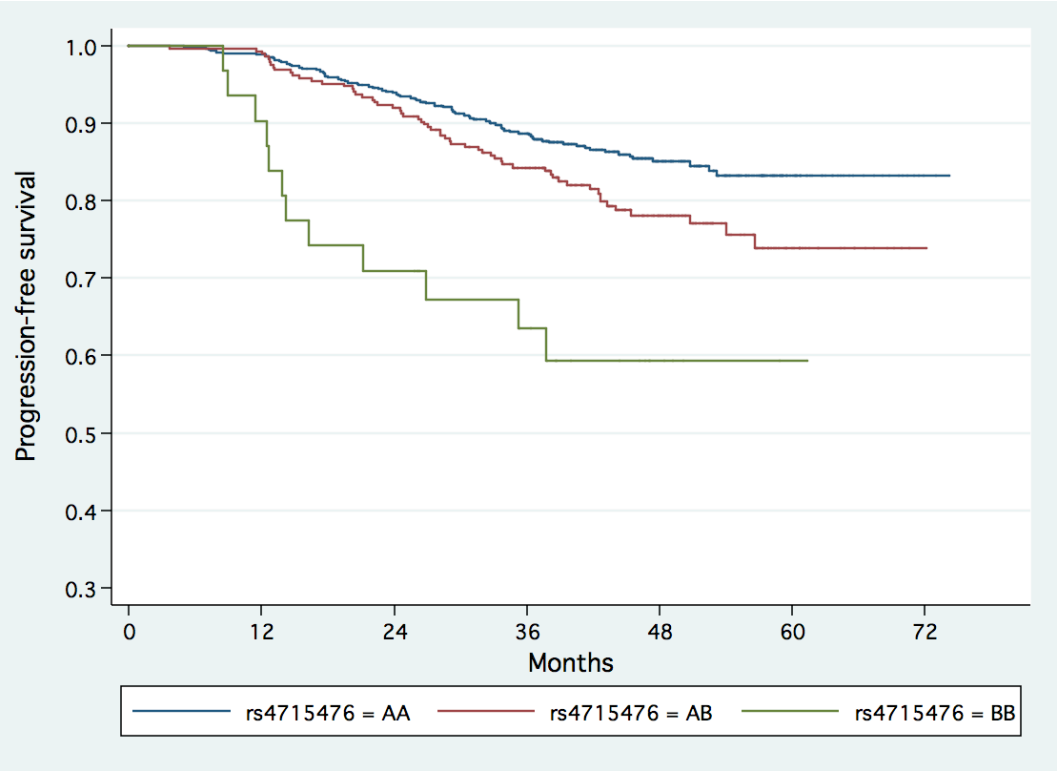


Figure 10-32 Kaplan-Meier curve for rs4776494

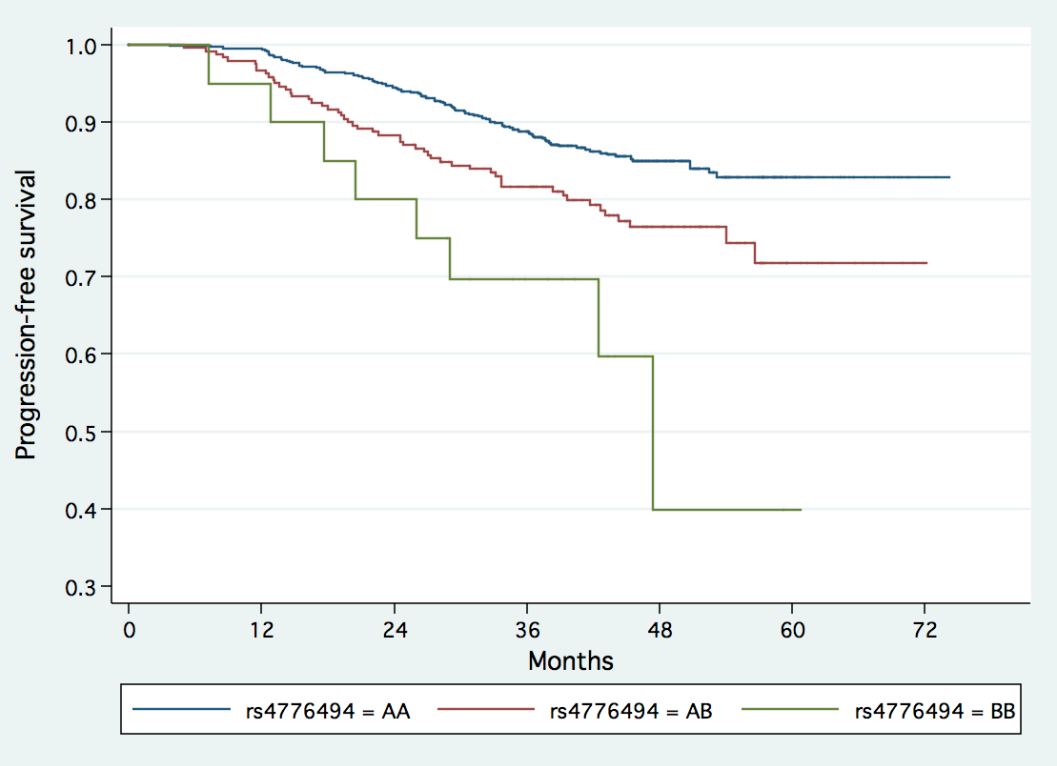


Figure 10-33 Kaplan-Meier curve for rs4978394

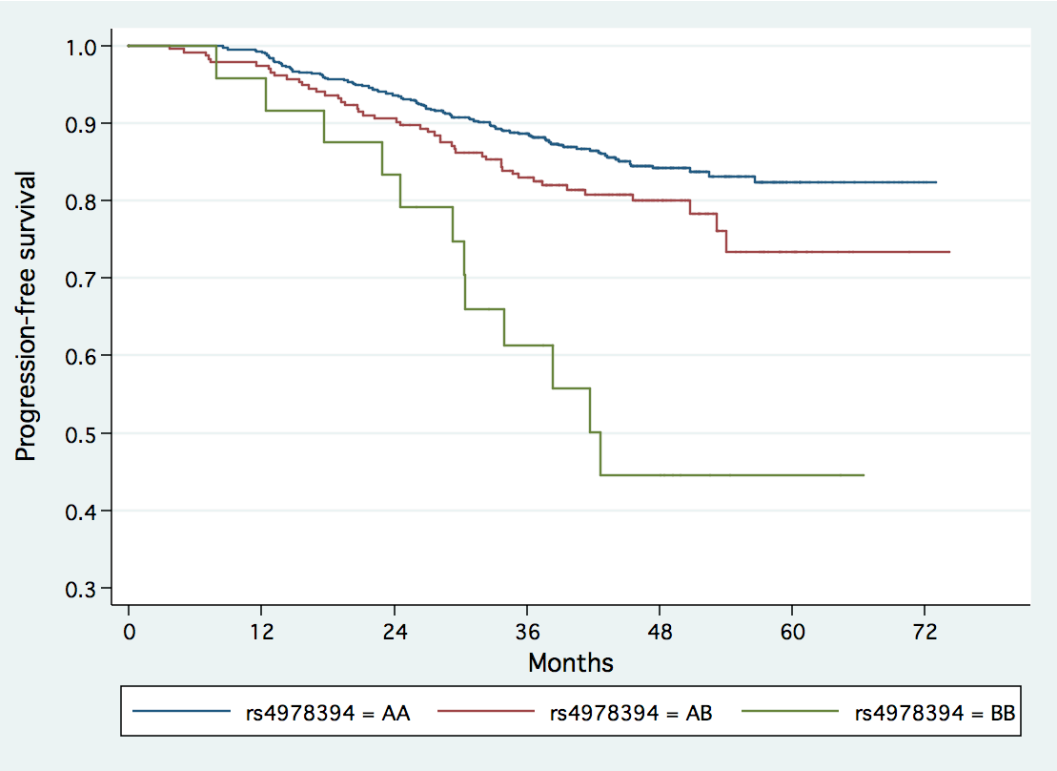


Figure 10-34 Kaplan-Meier curve for rs5997921

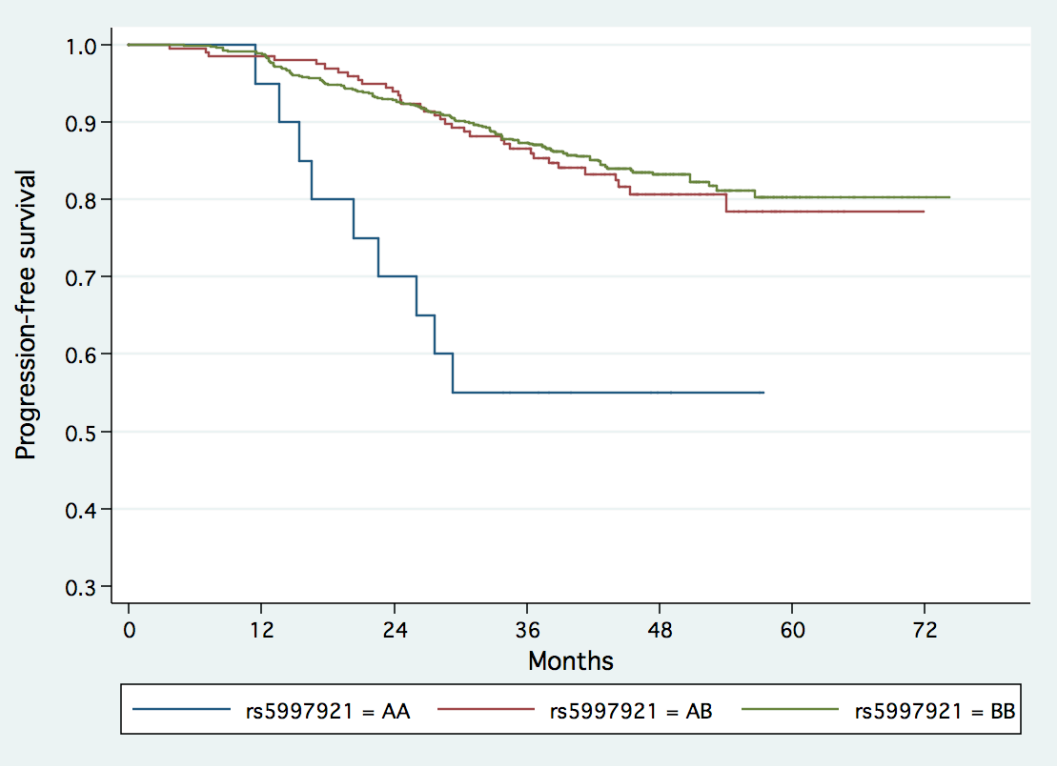


Figure 10-35 Kaplan-Meier curve for rs6518956

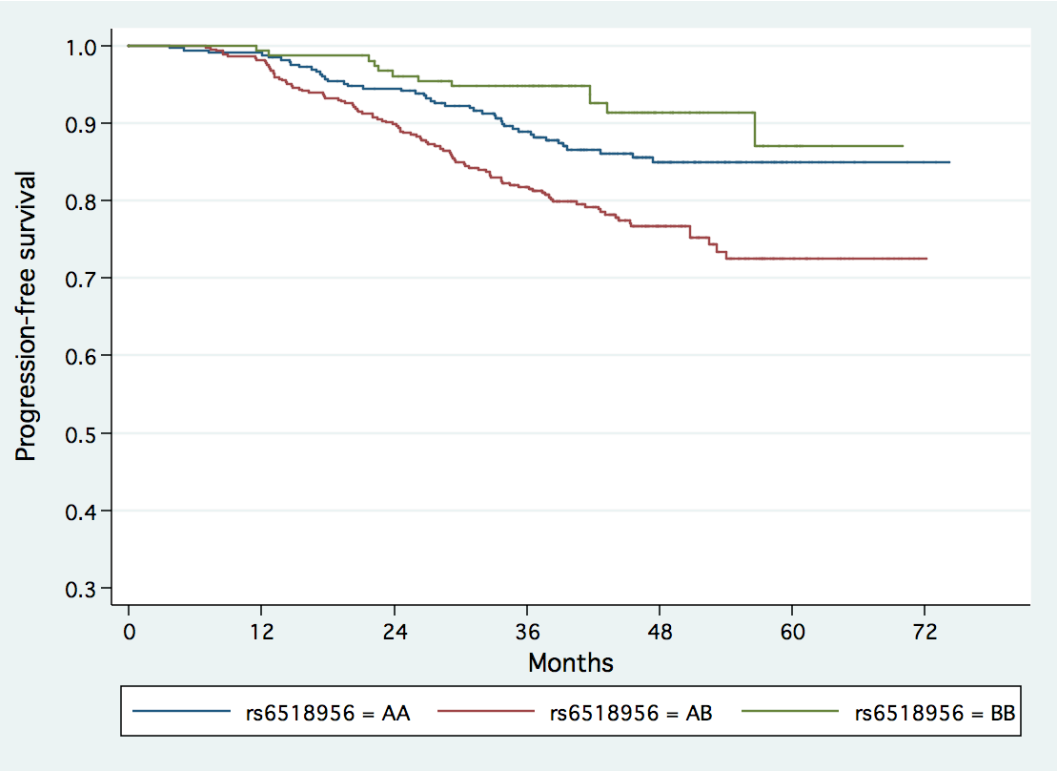


Figure 10-36 Kaplan-Meier curve for rs6972789

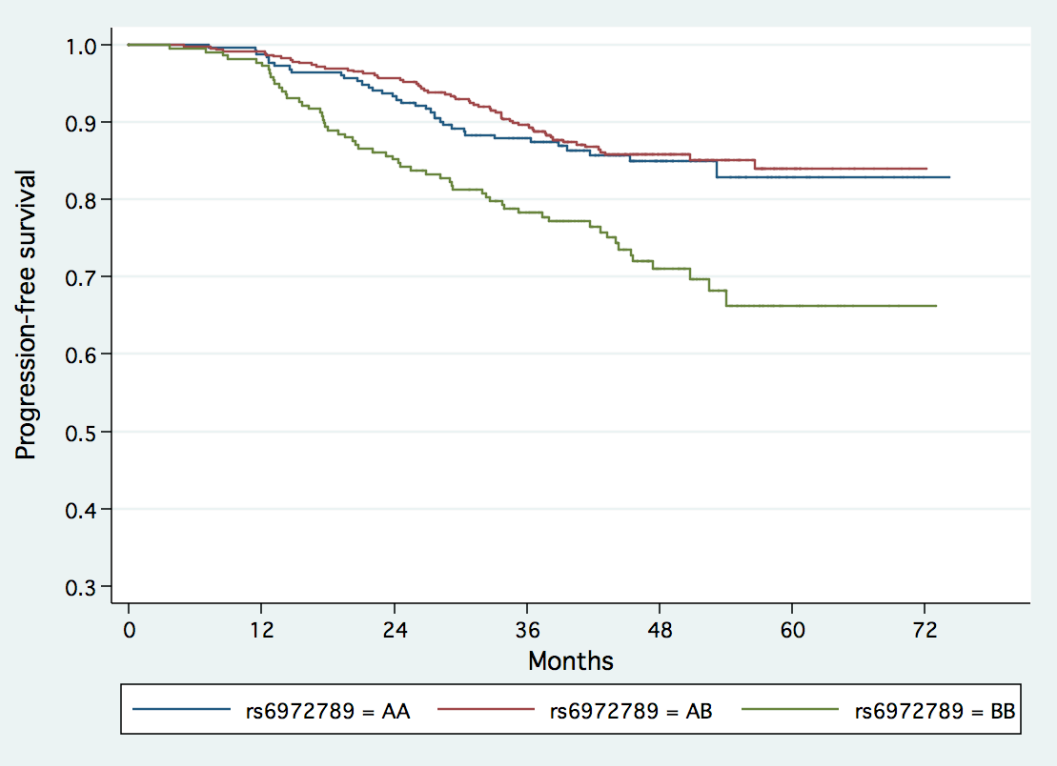


Figure 10-37 Kaplan-Meier curve for rs7007146

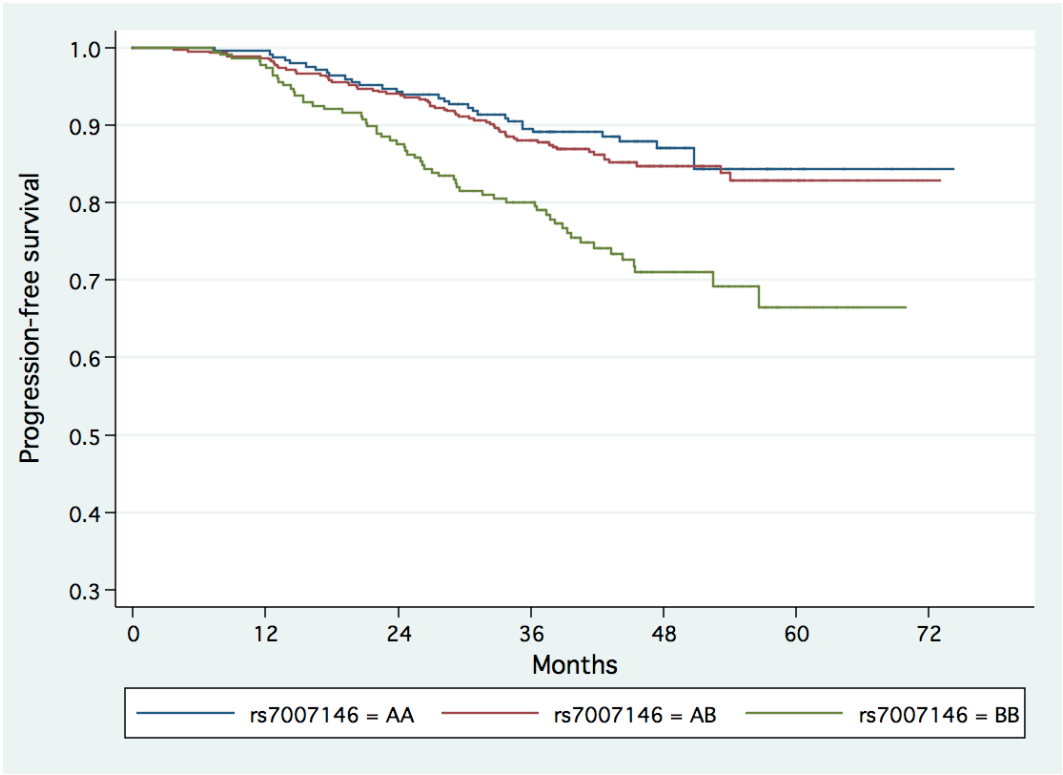


Figure 10-38 Kaplan-Meier curve for rs7556894

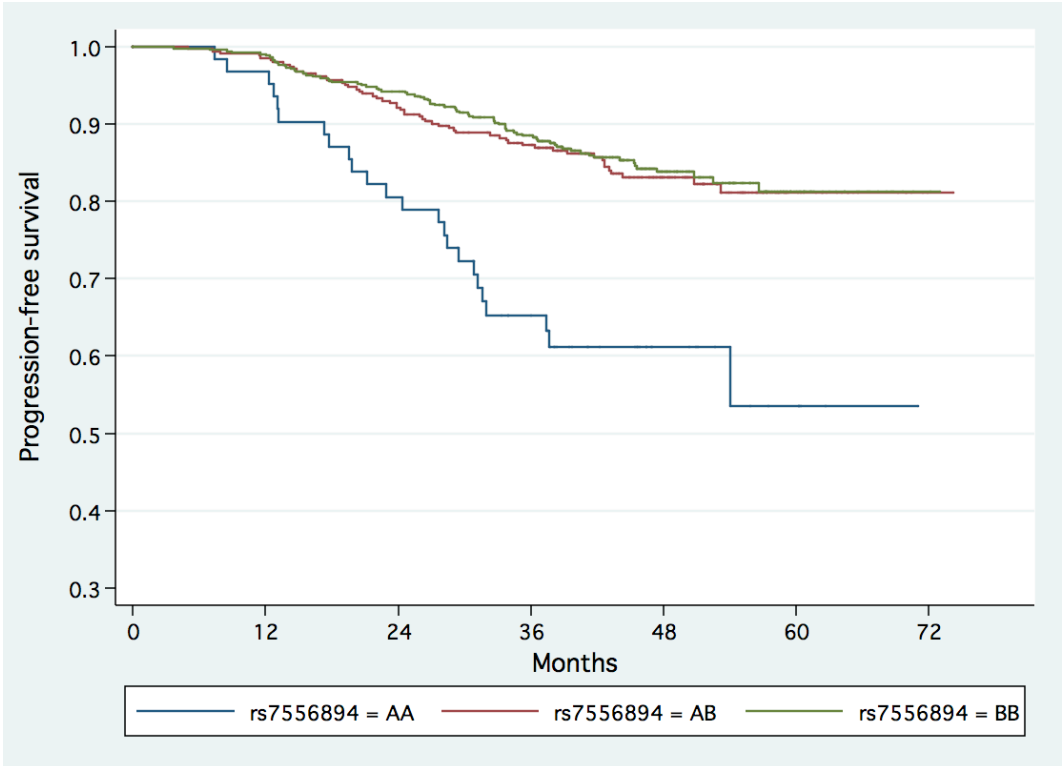


Figure 10-39 Kaplan-Meier curve for rs7600624

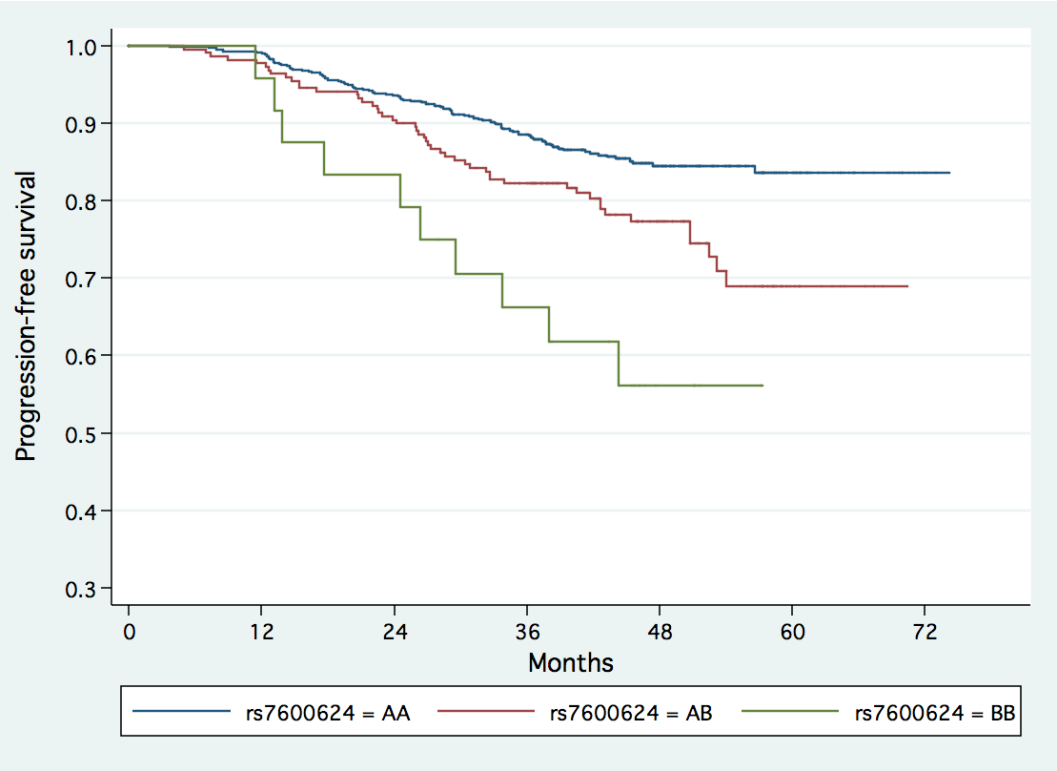


Figure 10-40 Kaplan-Meier curve for rs7866165

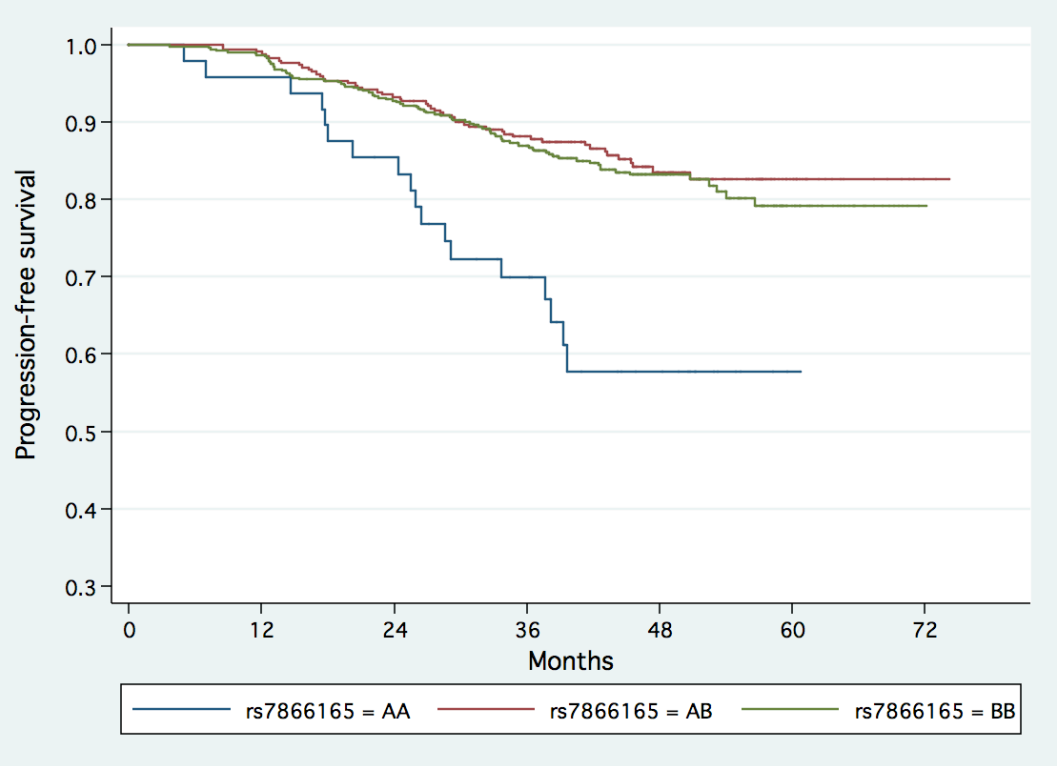


Figure 10-41 Kaplan-Meier curve for rs7912136

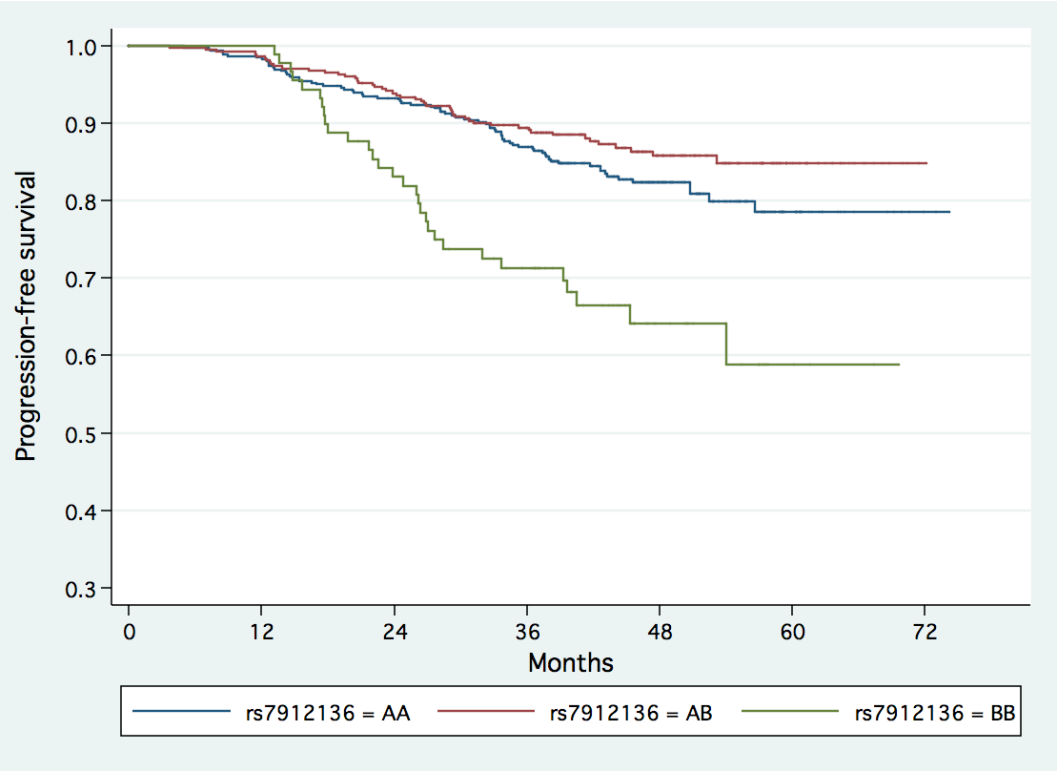


Figure 10-42 Kaplan-Meier curve for rs9315425

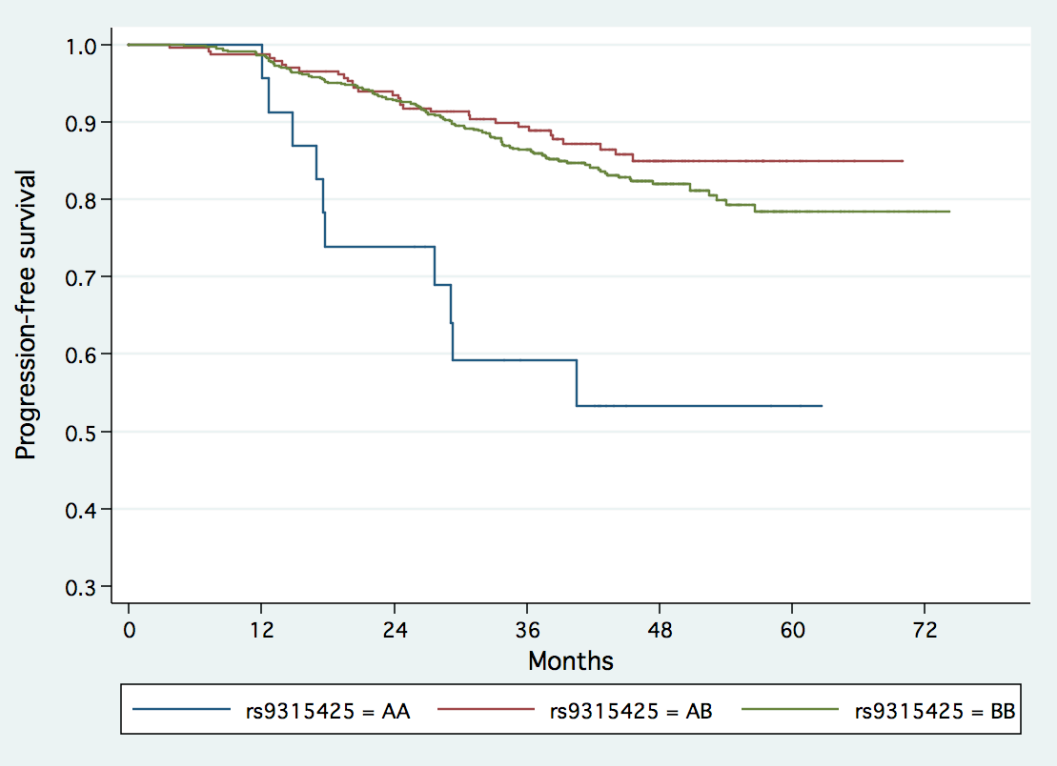


Figure 10-43 Kaplan-Meier curve for rs9514816

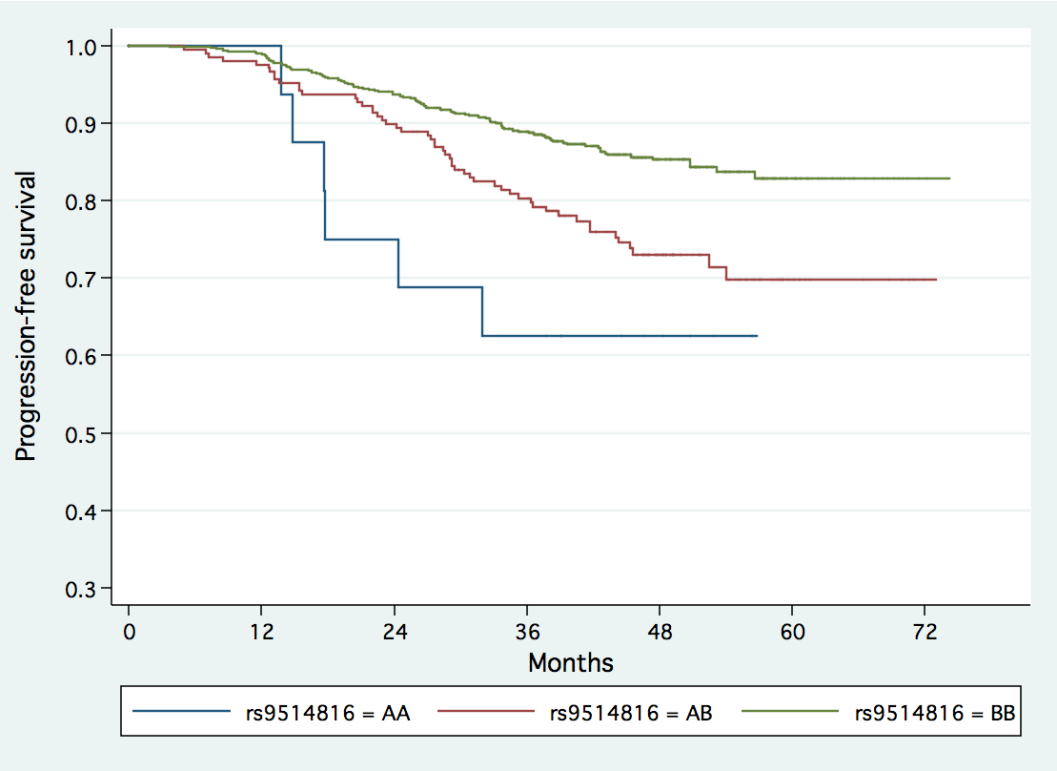


Figure 10-44 Kaplan-Meier curve for rs9533457

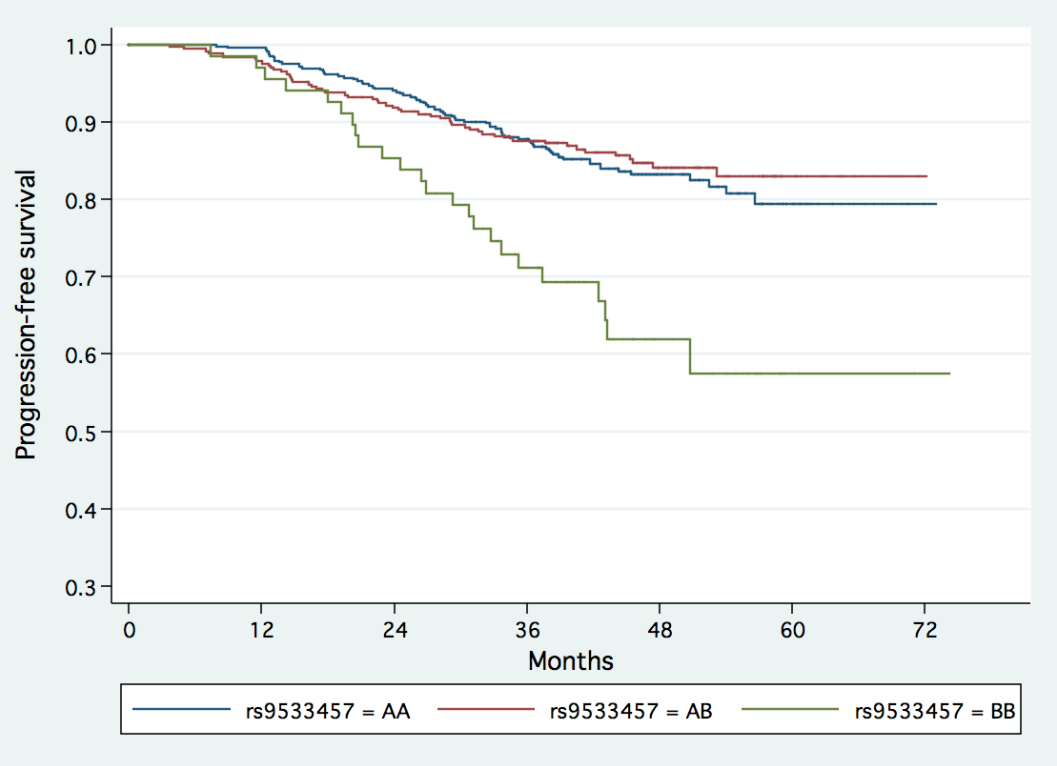


Figure 10-45 Kaplan-Meier curve for rs10429965

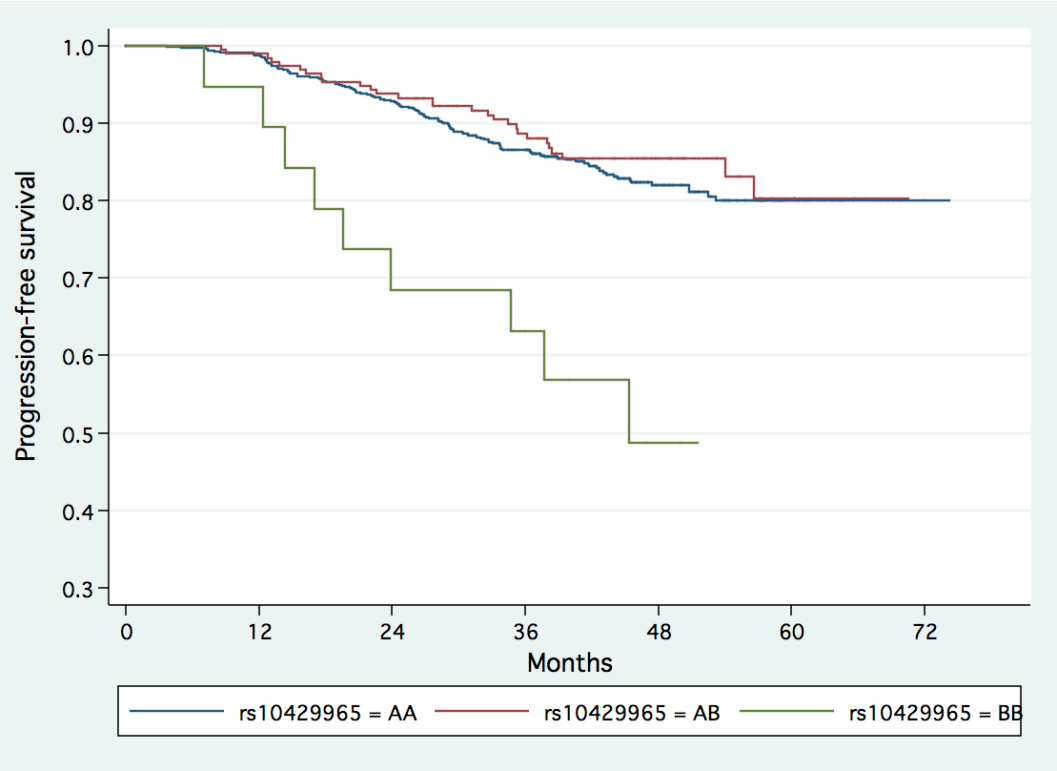


Figure 10-46 Kaplan-Meier curve for rs10510044

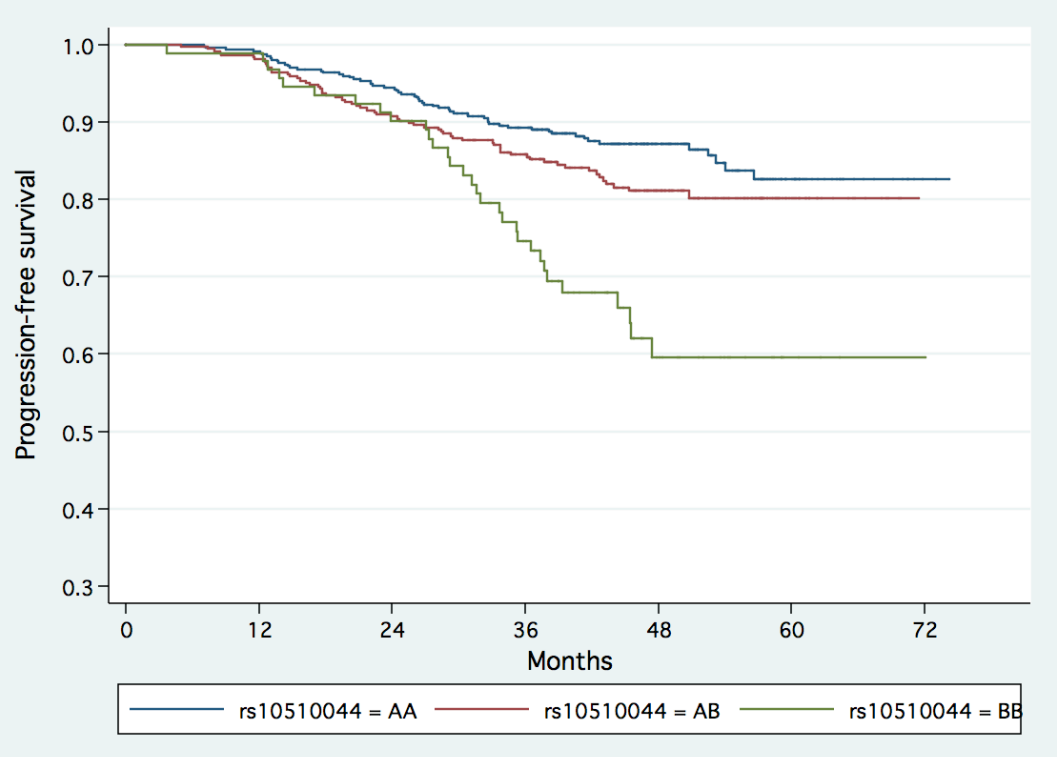


Figure 10-47 Kaplan-Meier curve for rs10842099

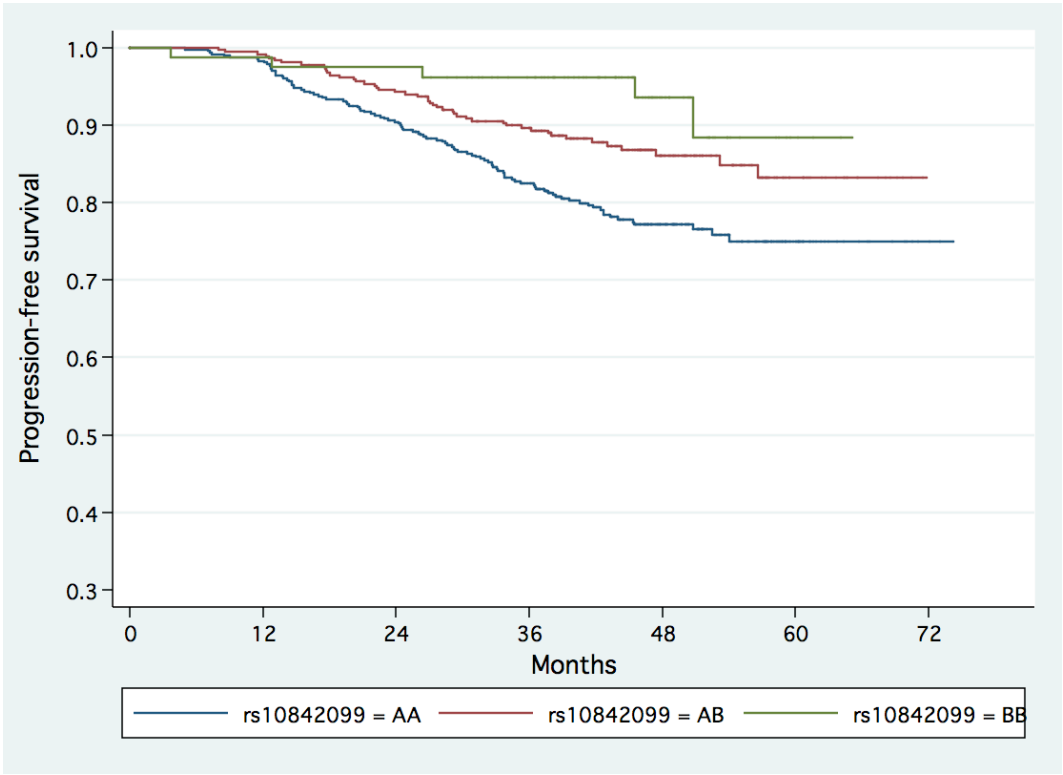
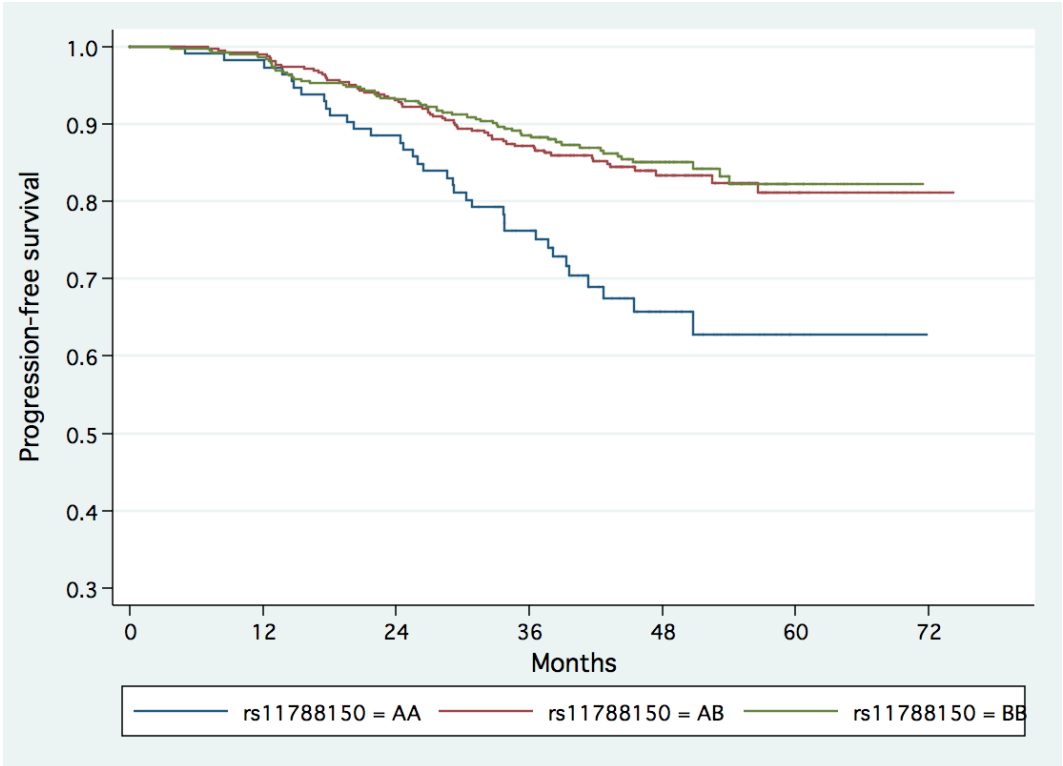


Figure 10-48 Kaplan-Meier curve for rs11788150



Manhattan plots by chromosome

Figure 10-49 Manhattan plot for chromosome 1

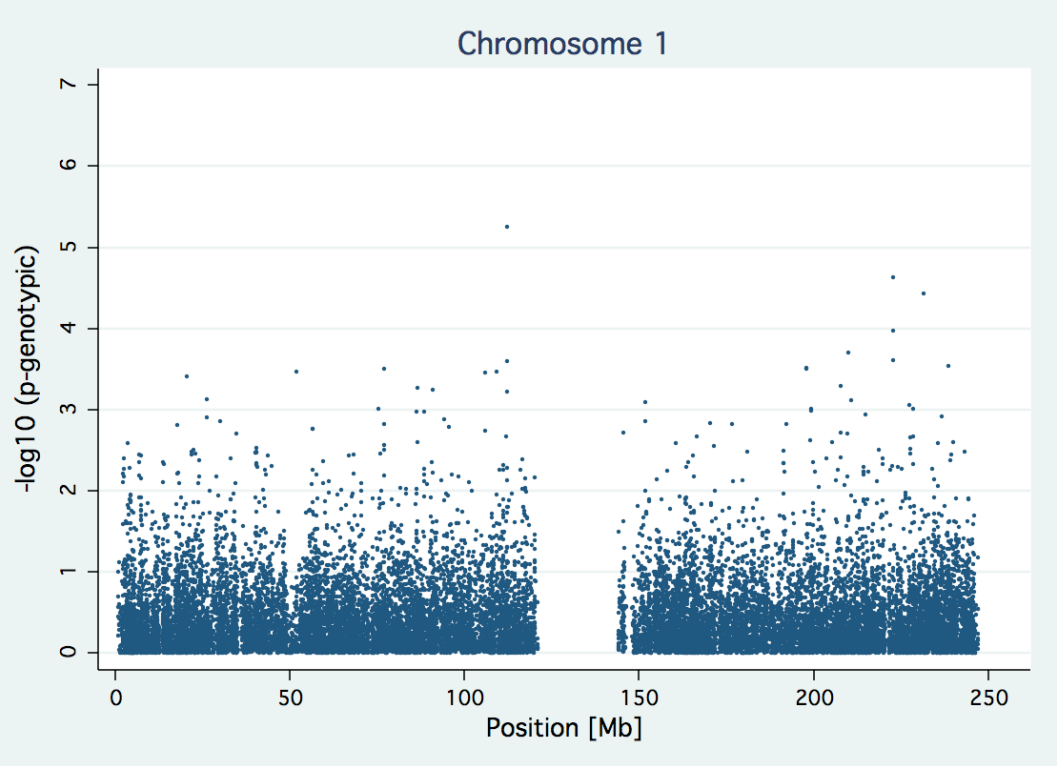


Figure 10-50 Manhattan plot for chromosome 2

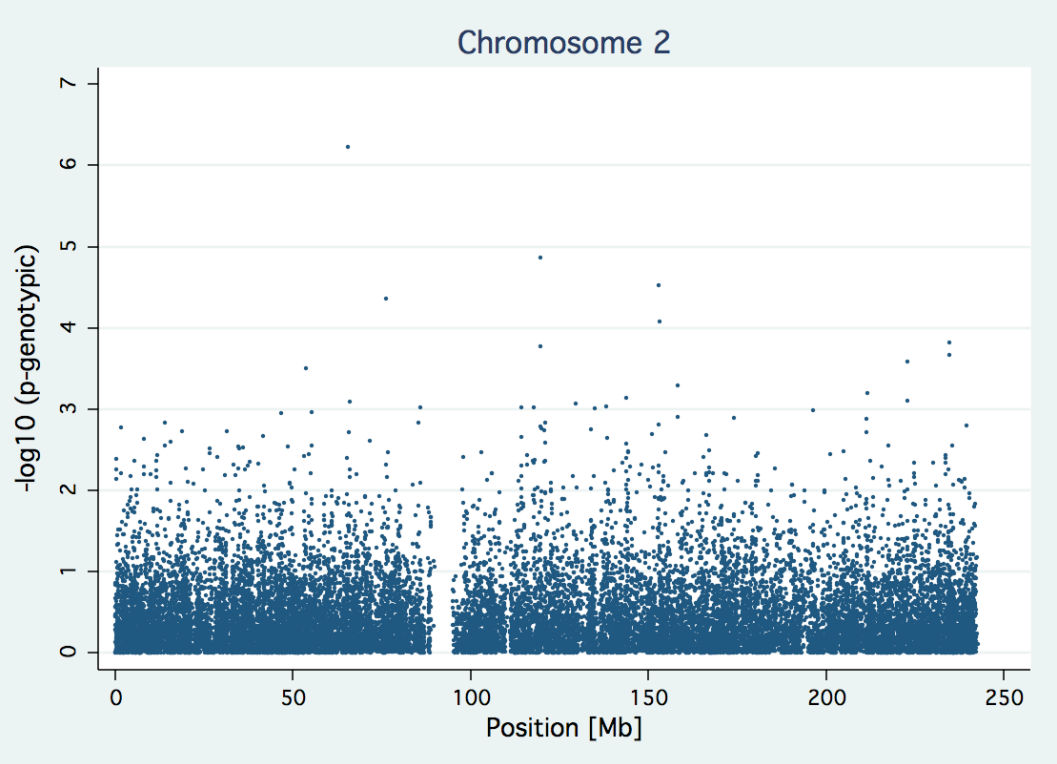


Figure 10-51 Manhattan plot for chromosome 3

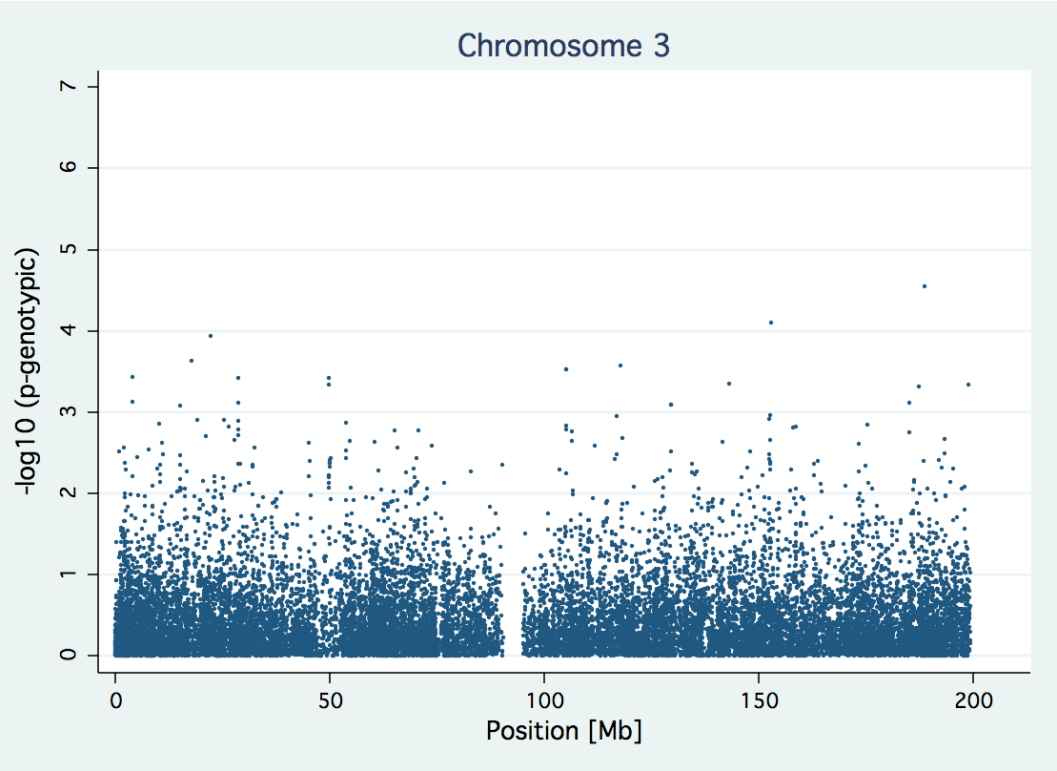


Figure 10-52 Manhattan plot for chromosome 4

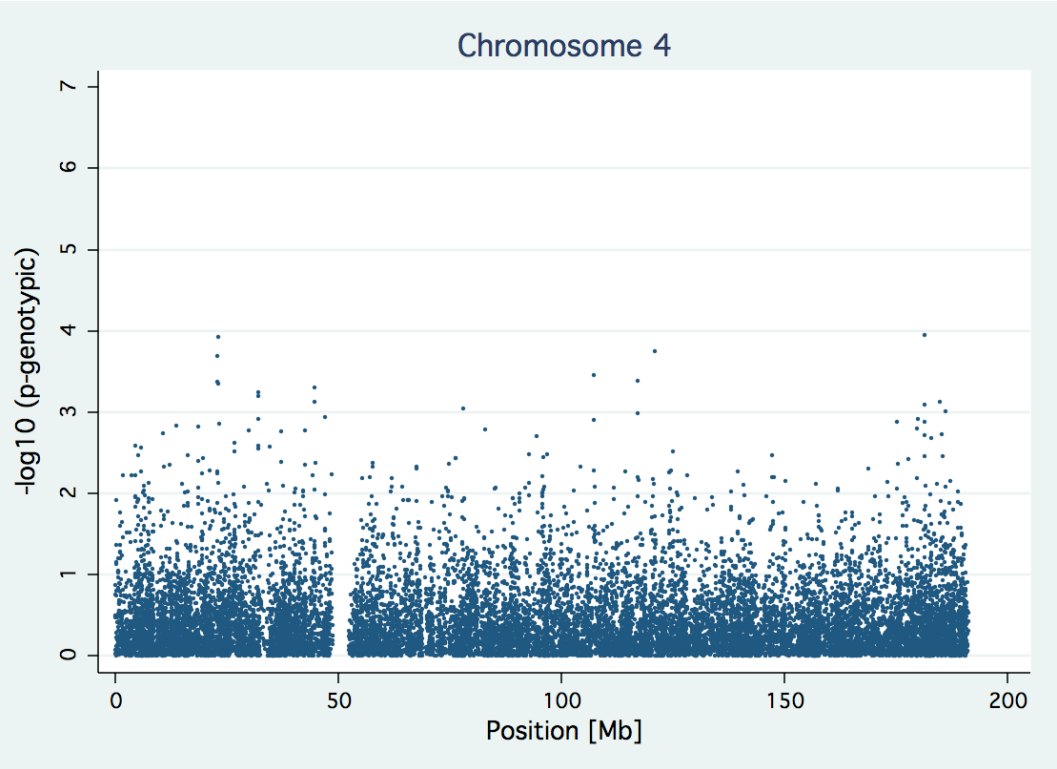


Figure 10-53 Manhattan plot for chromosome5

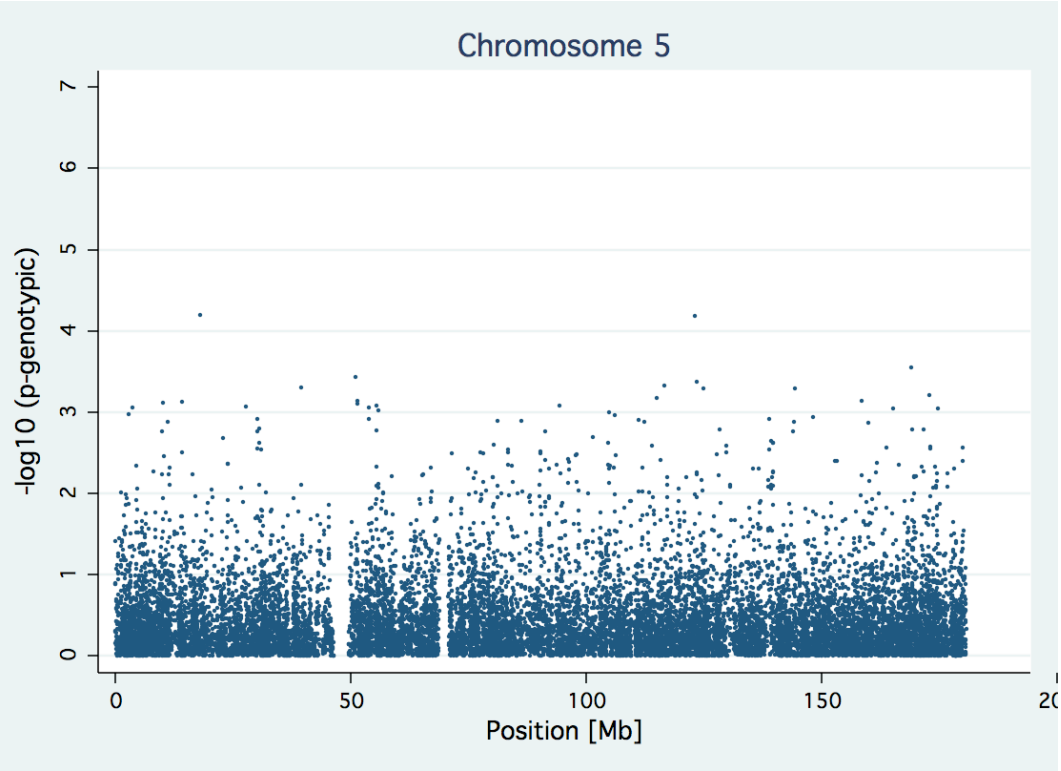


Figure 10-54 Manhattan plot for chromosome 6

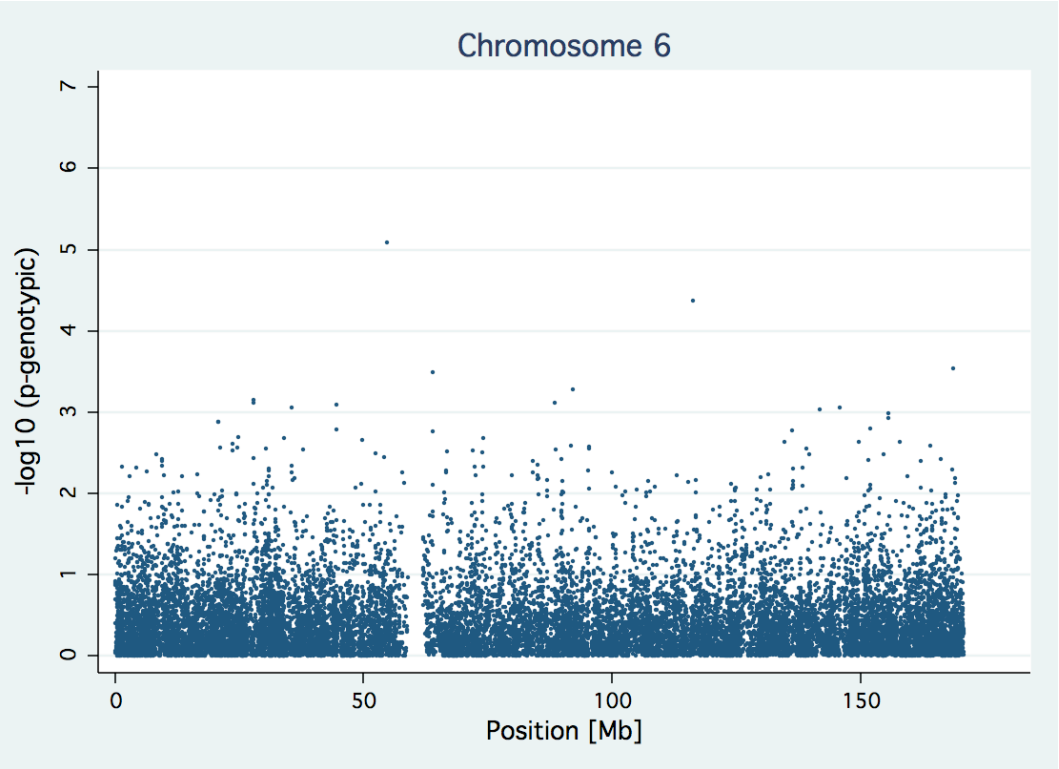


Figure 10-55 Manhattan plot for chromosome 7

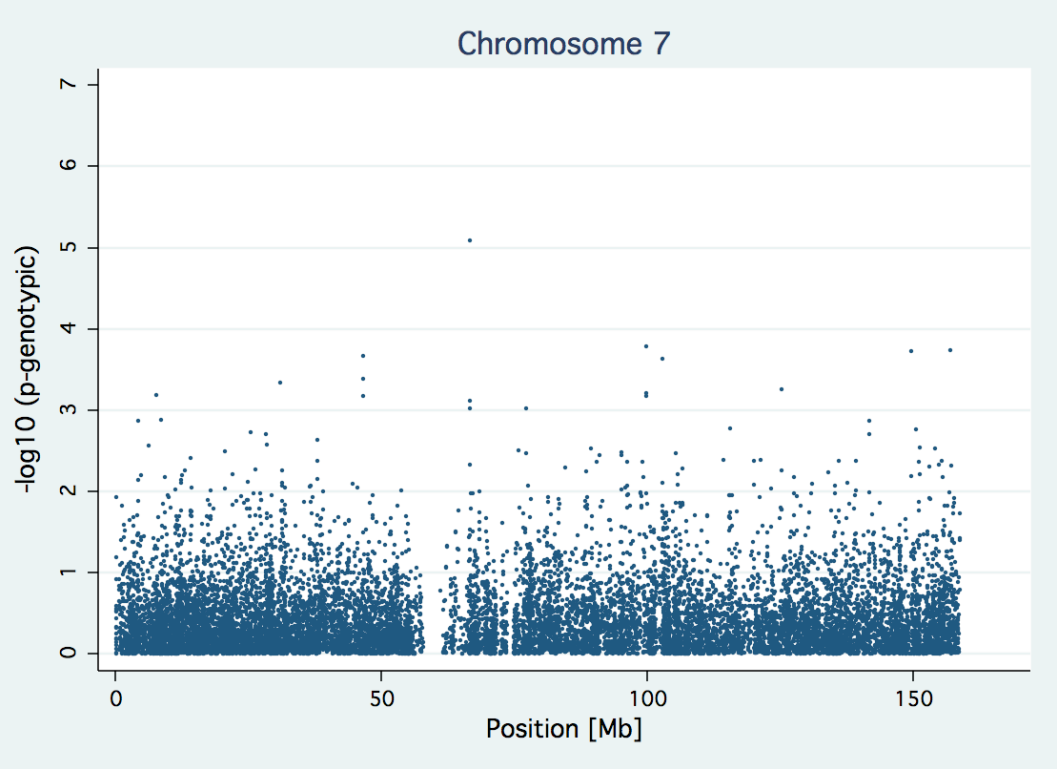


Figure 10-56 Manhattan plot for chromosome 8

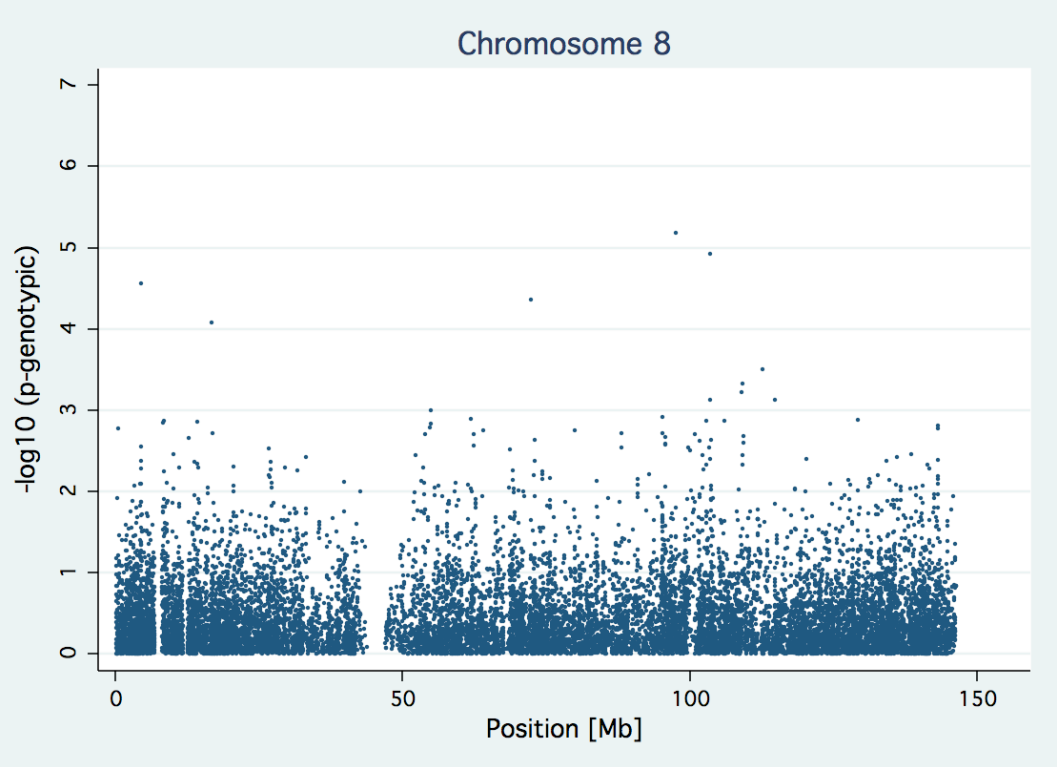


Figure 10-57 Manhattan plot for chromosome 9

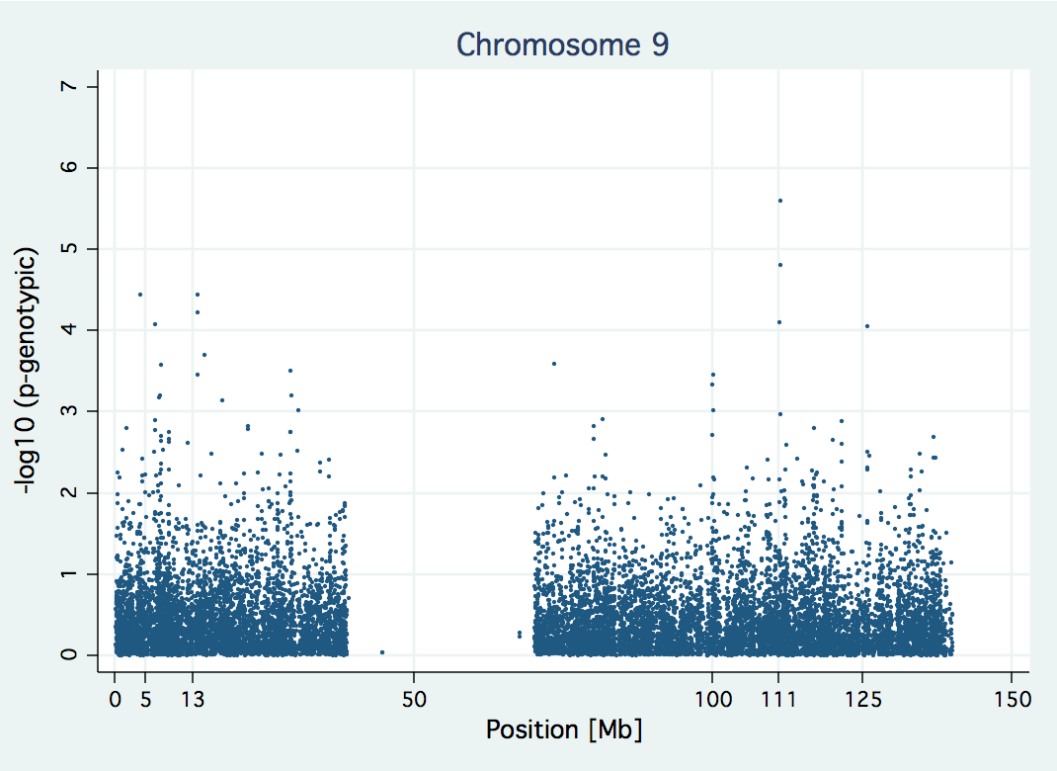


Figure 10-58 Manhattan plot for chromosome 10

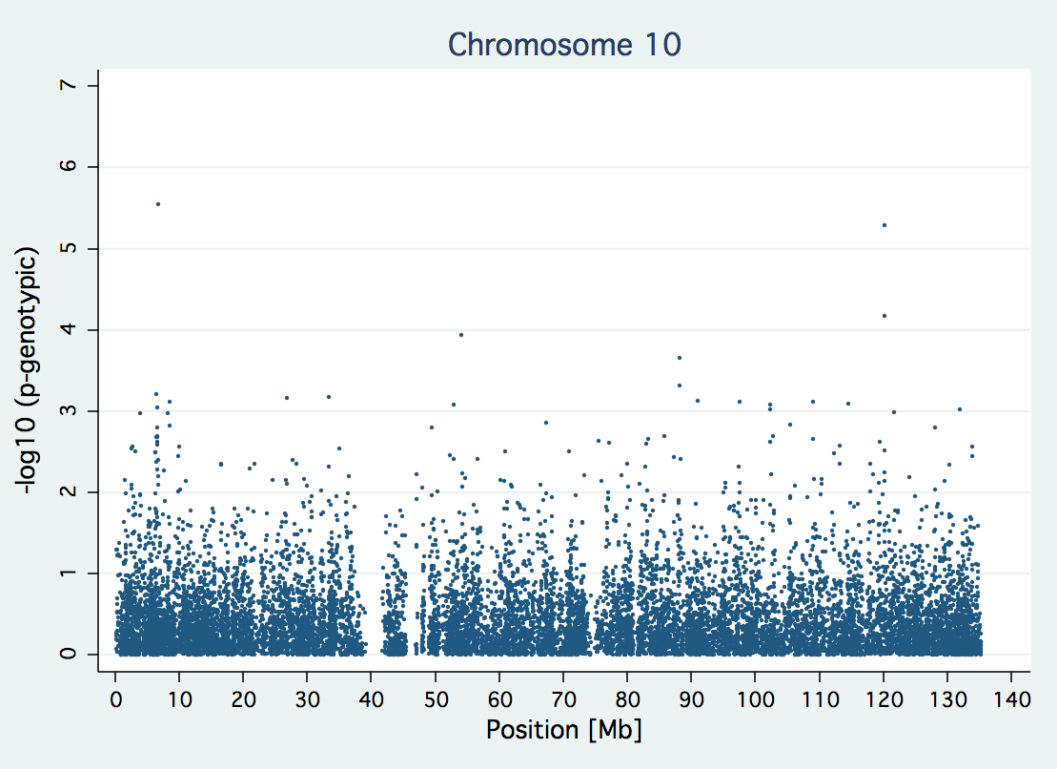


Figure 10-59 Manhattan plot for chromosome 11

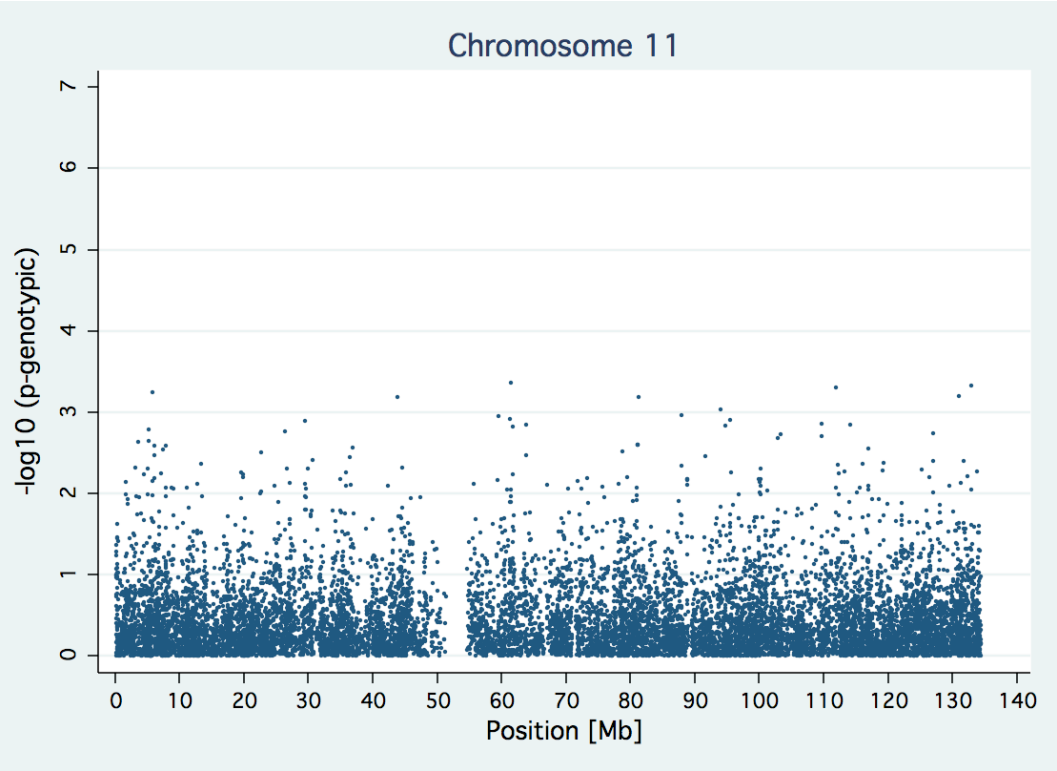


Figure 10-60 Manhattan plot for chromosome 12

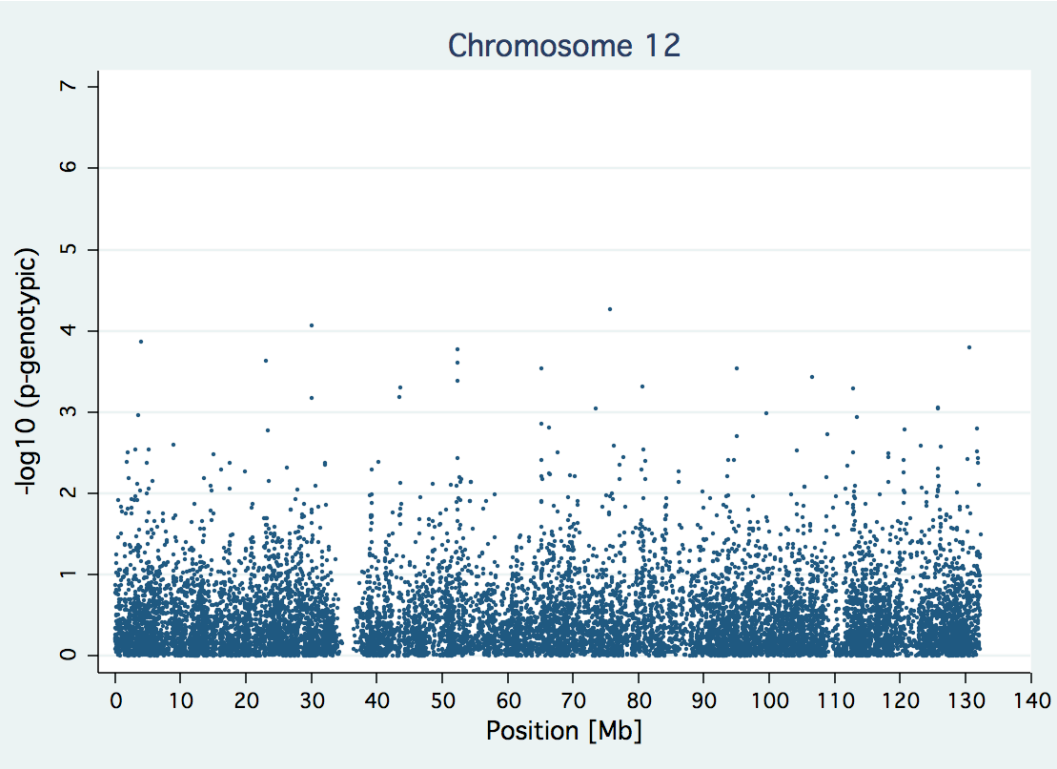


Figure 10-61 Manhattan plot for chromosome 13

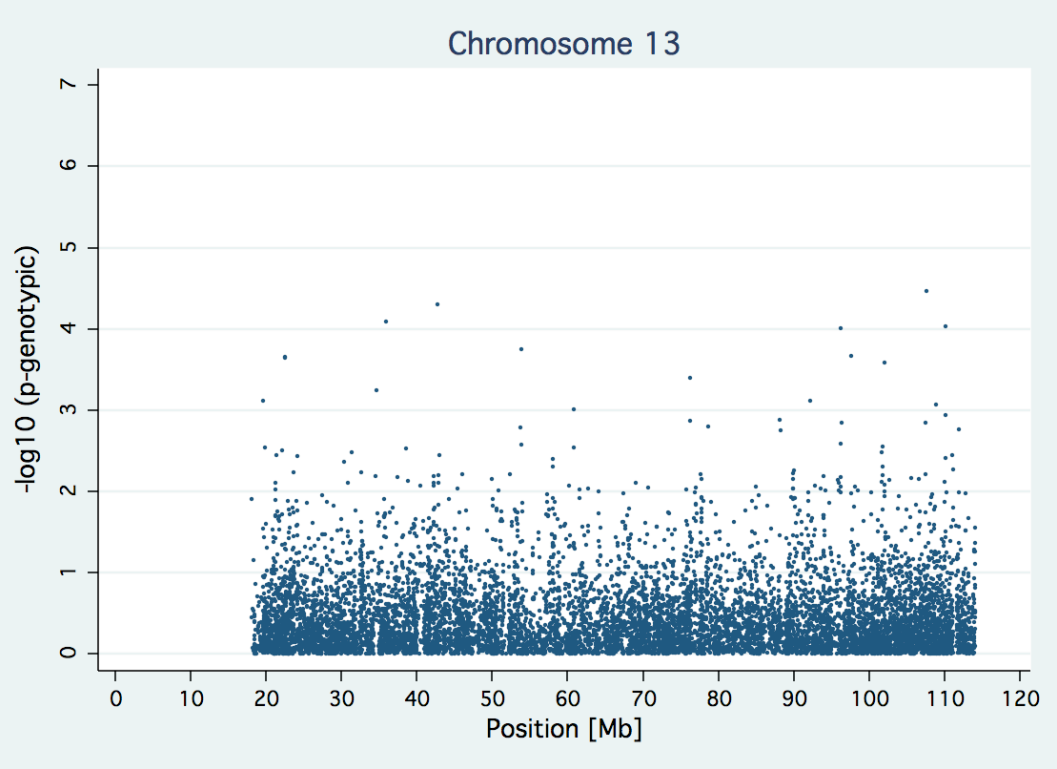


Figure 10-62 Manhattan plot for chromosome 14

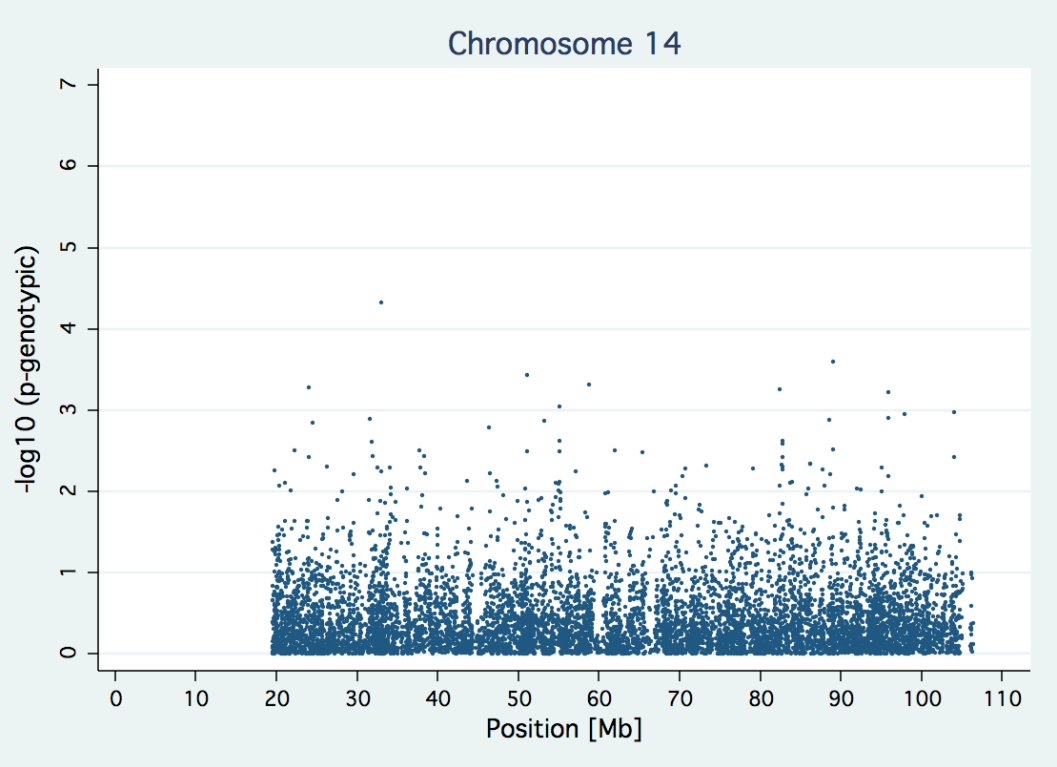


Figure 10-63 Manhattan plot for chromosome 15

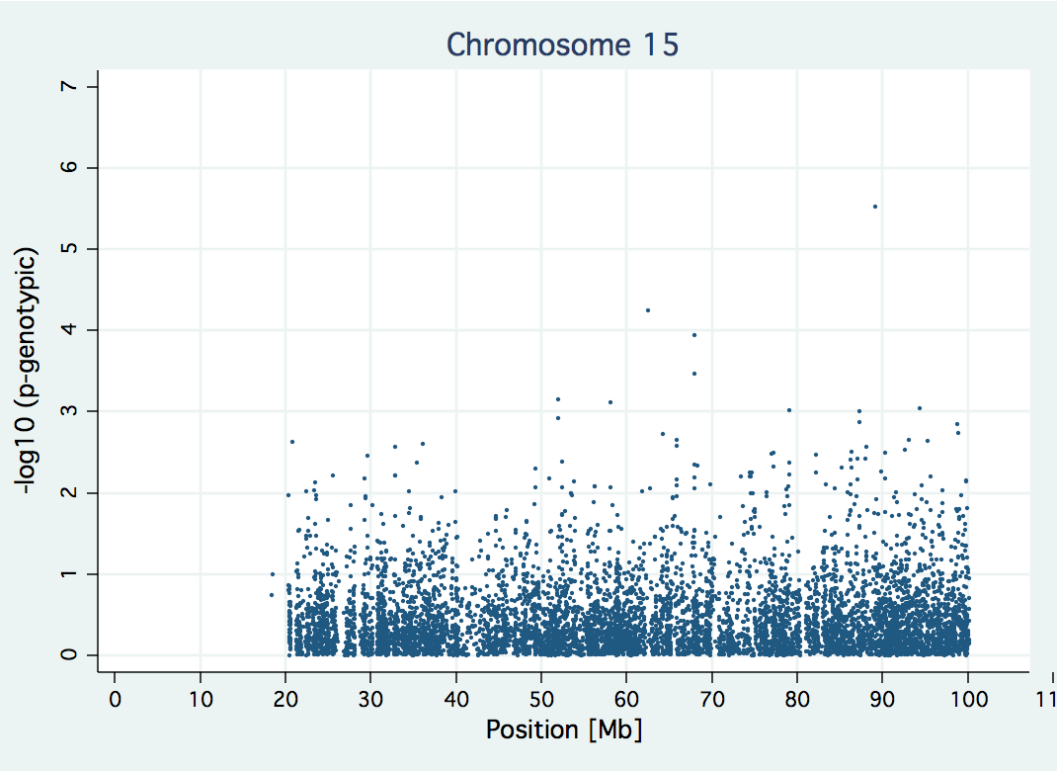


Figure 10-64 Manhattan plot for chromosome 16

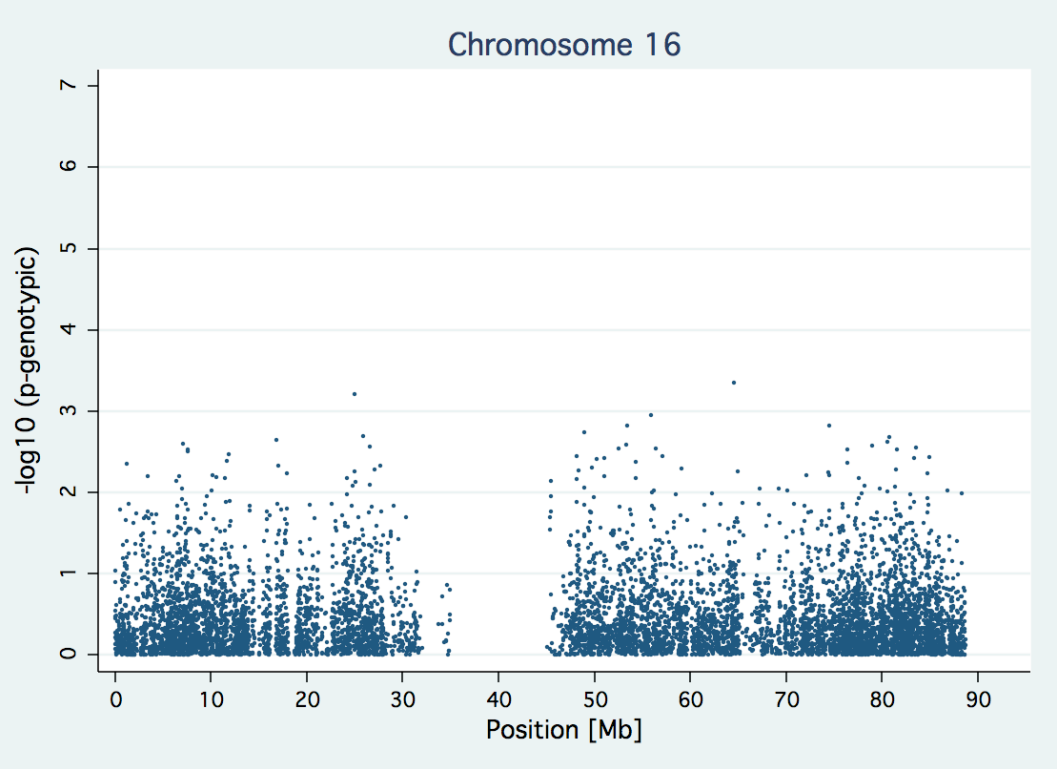


Figure 10-65 Manhattan plot for chromosome 17

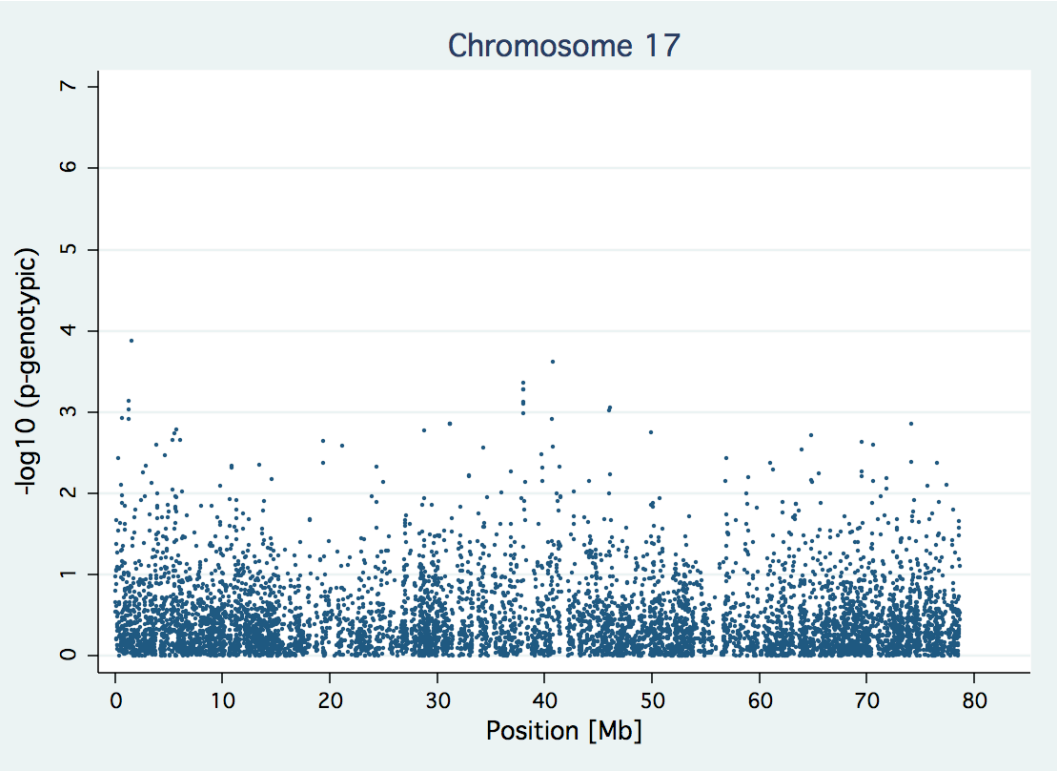


Figure 10-66 Manhattan plot for chromosome 18

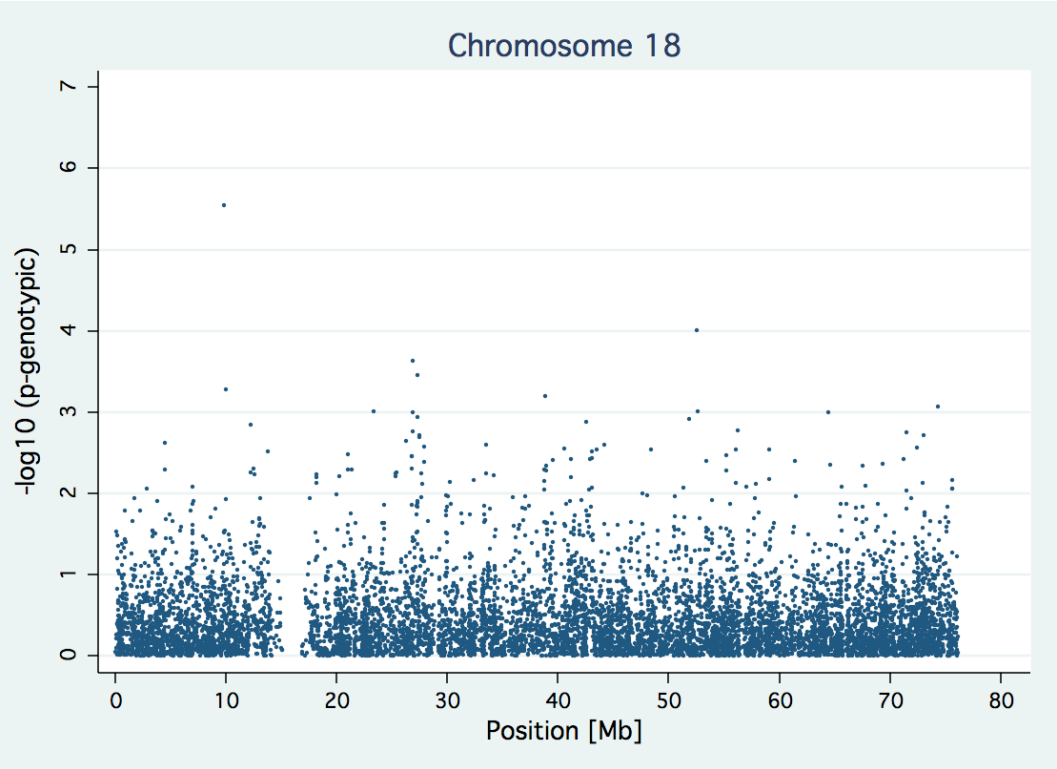


Figure 10-67 Manhattan plot for chromosome 19

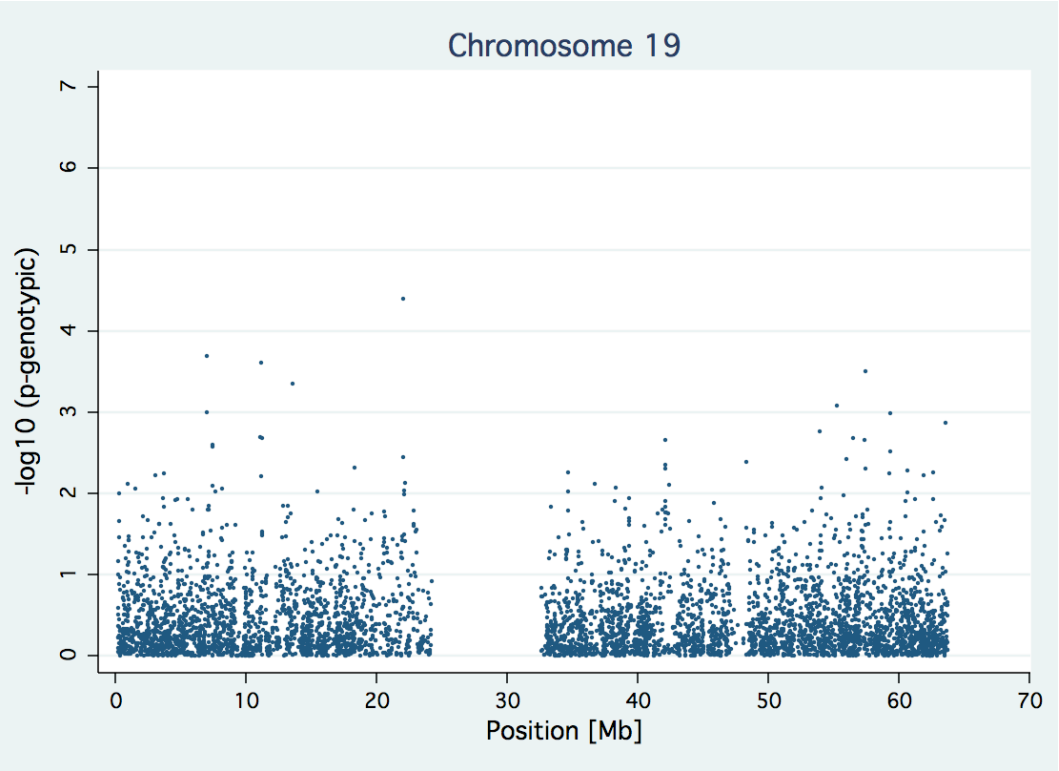


Figure 10-68 Manhattan plot for chromosome 20

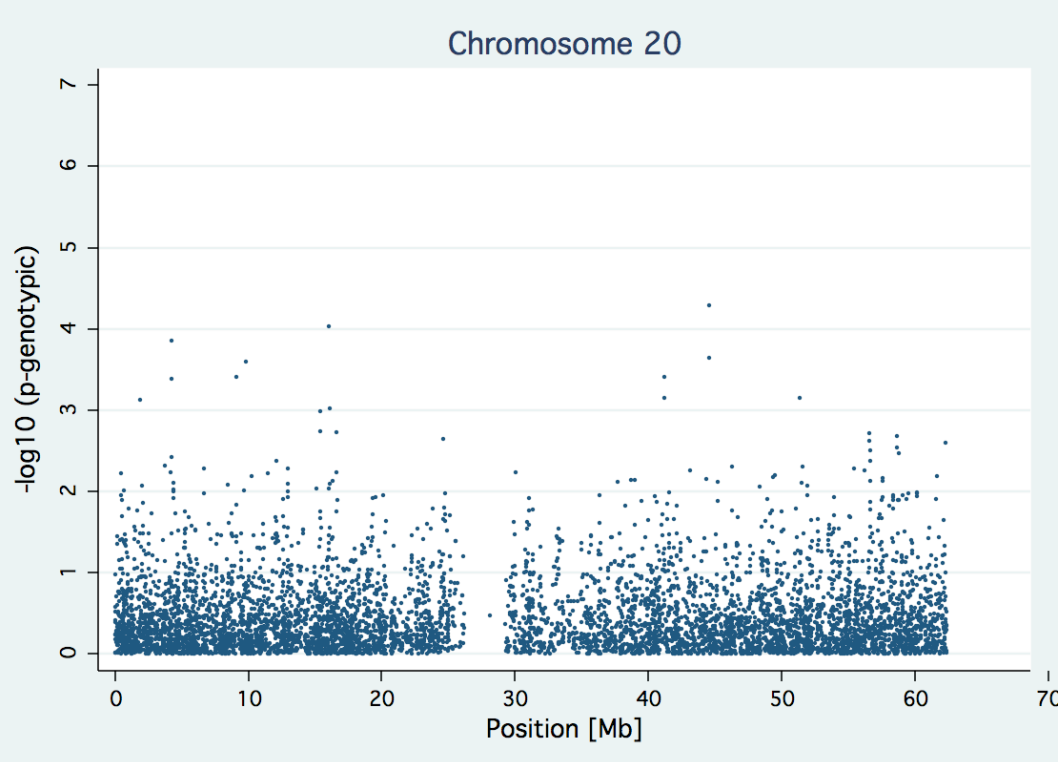


Figure 10-69 Manhattan plot for chromosome 21

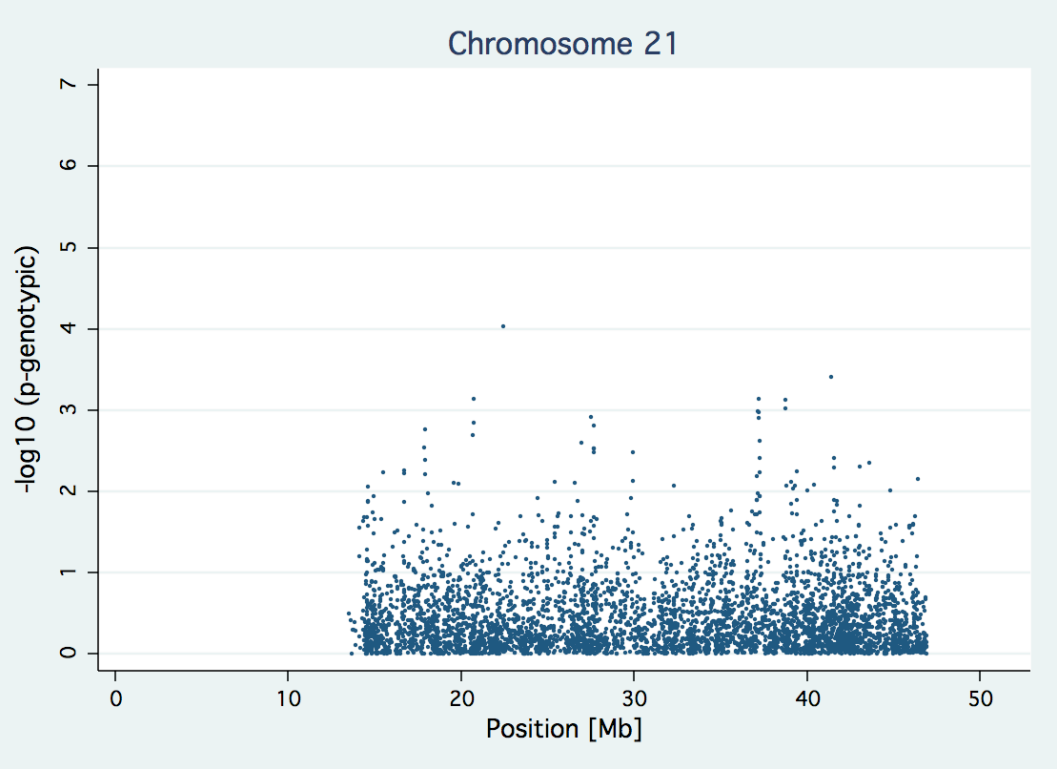
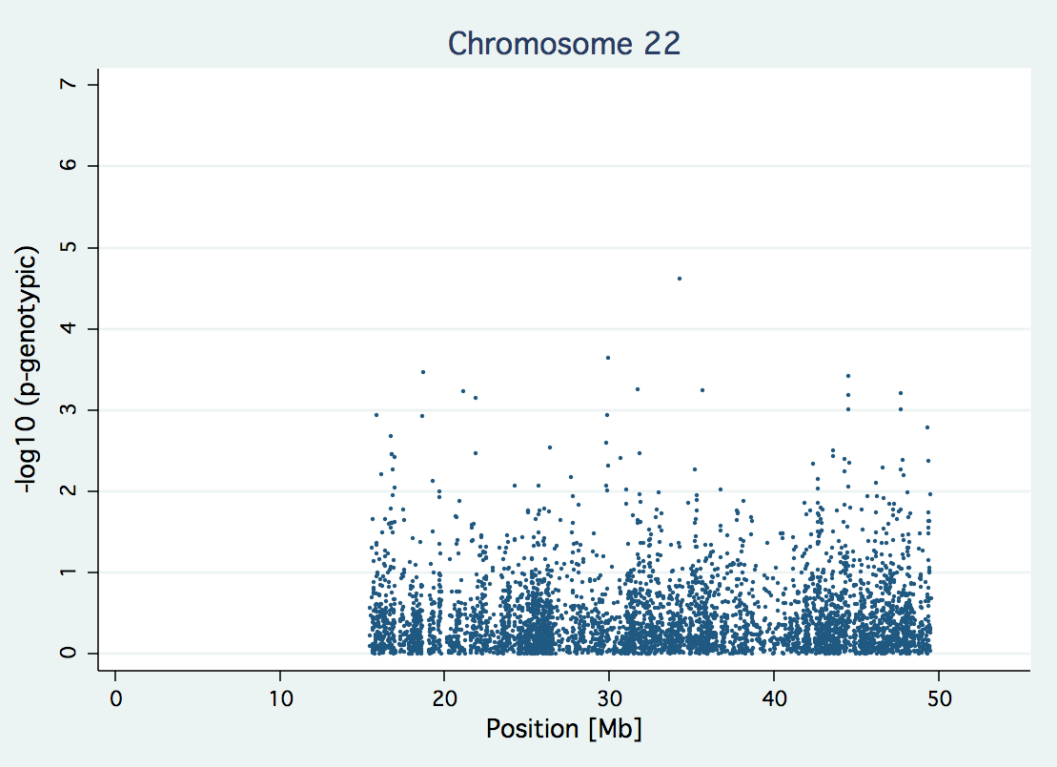


Figure 10-70 Manhattan plot for chromosome 22



Appendix C

Chapter 4

Table 10-8 Primers for KASPar genotyping

Locus	Type	Primer
rs1034116	Allele 1	GAAGGTGACCAAGTTCATGCTCAGGGGCTGTCGAGGCA
	Allele 2	GAAGGTCGGAGTCAACGGATTCAGGGGCTGTCGAGGCG
	Common	GCCATTCTGCAAAGGGGCAAAATGTT
rs1350308	Allele 1	GAAGGTGACCAAGTTCATGCTGTGTGAGTATTAAGTCATTTGCTAATTGTT`
	Allele 2	GAAGGTCGGAGTCAACGGATTGTGAGTATTAAGTCATTTGCTAATTGTG
	Common	GGTTGGAGATTTAGCATTTGTATTTGTGAA
rs1571583	Allele 1	GAAGGTGACCAAGTTCATGCTAAACCGTCACATCTCCCATCTTCTA
	Allele 2	GAAGGTCGGAGTCAACGGATTCCGTCACATCTCCCATCTTCTG
	Common	TAGAATAGTCTAGTAAAAACACTCCCACTT
rs1822917	Allele 1	GAAGGTGACCAAGTTCATGCTGTTCTTAGACTCTGCAGTTACTTATTCT
	Allele 2	GAAGGTCGGAGTCAACGGATTCTTAGACTCTGCAGTTACTTATTCC
	Common	GAACATGAGTAGTGGCTTTTGTACCTTTT
rs1924597	Allele 1	GAAGGTGACCAAGTTCATGCTTTTGATTGGTATTTCACTAAATTCATAGATCA
	Allele 2	GAAGGTCGGAGTCAACGGATTGATTGGTATTTCACTAAATTCATAGATCG
	Common	ATATAGTGAATATACCAATTCTCCCCAAGT
rs2589183	Allele 1	GAAGGTGACCAAGTTCATGCTTGAACCAGGTCCAAATTGCTGAAT
	Allele 2	GAAGGTCGGAGTCAACGGATTGAACCAGGTCCAAATTGCTGAAC
	Common	GGTCTTCTTACCGTACCTGCACTTA
rs3784780	Allele 1	GAAGGTGACCAAGTTCATGCTGTTTTAAAGATTAATGACAATGAAGGGGT
	Allele 2	GAAGGTCGGAGTCAACGGATTGTTTTAAAGATTAATGACAATGAAGGGGC
	Common	GGTAGCCATTGTTCAACCATTTGTCATAT
rs4649314	Allele 1	GAAGGTGACCAAGTTCATGCTCTATCTGGGCAGTTACGTCT
	Allele 2	GAAGGTCGGAGTCAACGGATTCTCTATCTGGGCAGTTACGTCG
	Common	GTGGAGAGTCAGAGAAGTTACTGCTA
rs6518956	Allele 1	GAAGGTGACCAAGTTCATGCTGAGGCTCCCATAGAAAGCTCTGA
	Allele 2	GAAGGTCGGAGTCAACGGATTAGGCTCCCATAGAAAGCTCTGG
	Common	CAAGGGTTATGGTTCTGGGGCAAA
rs7556894	Allele 1	GAAGGTGACCAAGTTCATGCTACAGCTTGCTTATGAATTGCAAGGAT
	Allele 2	GAAGGTCGGAGTCAACGGATTCAGCTTGCTTATGAATTGCAAGGAC
	Common	CGTTTCTTCATTGCCGAGTTACCATATTT
rs7600624	Allele 1	GAAGGTGACCAAGTTCATGCTGAAGTGCTCATACCAGAGAGAT
	Allele 2	GAAGGTCGGAGTCAACGGATTGAAGTGCTCATACCAGAGAGAC
	Common	GTTCTTGGCGGTTTCCAAACGACTT

Primers for ten SNPs typed in the Danish cohort, and one SNP typed in the Finnish cohort (rs4649314).

Table 10-9 PCR conditions

SNP	[$Mg^{2=}$]	Programme
rs1034116	2.6mM	Standard
rs1350308	3.4mM	Standard
rs1571583	2.6mM	Standard
rs1822917	2.6mM	Standard
rs1924597	2.6mM	Standard
rs2589183	2.2mM	Standard
rs3784780	2.2mM	Standard
rs4649314	2.2mM	Standard
rs6518956	2.6mM	Standard
rs7556894	2.6mM	Standard
rs7600624	2.6mM	Standard

Conditions used for ten SNPs typed in Danish cohort, and one SNP typed in Finnish cohort (rs4649314).

Table 10-10 Primer sequences for KASPar PCR

Locus	Type	Primer
rs10411210	Allele 1	GAAGGTGACCAAGTTCATGCTAACAAGCACCAACGGTTTCCCG
	Allele 2	GAAGGTCGGAGTCAACGGATTCAACAAGCACCAACGGTTTCCCA
	Common	GCCAGAGCGGAGCTTGGCAAAA
rs10758669	Allele 1	GAAGGTGACCAAGTTCATGCTTGGTGGGAAATATAAAATCATGGAATT
	Allele 2	GAAGGTCGGAGTCAACGGATTCTTGGTGGGAAATATAAAATCATGGAATC
	Common	GCAGCAGCAGAAAGAGAAAAAGTTAGATT
rs16892766	Allele 1	GAAGGTGACCAAGTTCATGCTGACATAAGGCATAACCTTTAACAGCT
	Allele 2	GAAGGTCGGAGTCAACGGATTGACATAAGGCATAACCTTTAACAGCG
	Common	CAGAACGGTCAGACGCAAACAGTTT
rs6983267	Allele 1	GAAGGTGACCAAGTTCATGCTCATAAAAATTCTTTGTACTTTTCTCAGTGC
	Allele 2	GAAGGTCGGAGTCAACGGATTCACATAAAAATTCTTTGTACTTTTCTCAGTGA
	Common	CGTCCTTTGAGCTCAGCAGATGAAA
rs961253	Allele 1	GAAGGTGACCAAGTTCATGCTCAGCAACTTCAATTAATCTTTCTGAATT
	Allele 2	GAAGGTCGGAGTCAACGGATTCAGCAACTTCAATTAATCTTTCTGAATG
	Common	CAGATTGAAAGTGCATACCAAGTATTGAG

Table 10-11 PCR conditions

SNP	$[Mg^{2=}]$	Programme
rs10411210	2.2mM	Standard
rs10758669	3.6mM	Standard
rs16892766	4.0mM	Standard
rs6983267	2.2mM	Standard
rs961253	2.2mM	Standard

Chapter 7

Table 10-12 rs6687758 primers for KASPar assay

Allele	Primer
A	GAAGGTGACCAAGTTCATGCTAAAGAATGTGCATCTCTAGATTCCATATTT
B	GAAGGTCGGAGTCAACGGATTGAATGTGCATCTCTAGATTCCATATTC
Common	GAGGTGGGGAGCTATCTGATCATAT

Table 10-13 rs6687758 PCR conditions for KASPar

SNP	$[Mg^{2=}]$	Programme
rs6687758	2.2mM $Mg2^{+}$	Standard with an additional 10 cycles to improve dye uptake

Table 10-14 rs6687758 primers for sequencing

Direction	Primer
Forward	GGGTGCTTCTGAGGACAGAG
Reverse	CCCAGCAGGAATGCTTAAAA

Table 10-15 PCR conditions and primer for putative CRC risk variants

Locus	Mg^{2+}	Temp	Additive	Type	Primer
TGFβR1	1.5	55	DMSO	Forward	9. 6GAGGTTTGCTGGGGTGAG
				Reverse	10. AGCAGGAGCGAGCCAGAG
APC E1317Q	2.5	55	Q-solution	Forward	11. AGAAAAGATTGGAACTAGGTCAGC
				Reverse	12. TGCAGTCTGCTGGATTTGGT
rs28931588/9	1.5	55	DMSO	Forward	13. TTTGATGGAGTTGGACATGG
				Reverse	14. CAGGACTTGGGAGGTATCCA

Table 10-16 PCR conditions and primers for insertion-deletion polymorphisms

Locus	Mg^{2+}	Temp	Additive	Type	Primer
rs10667315	2.5	55	Q-solution	Forward	AGGGGAGACCACCTGTTTCT
				Reverse	TTTACTGCCCCATAGGCAAC
rs10697058	1.5	55	Q-solution	Forward	ATGACAAGGTCTGCCCACAT
				Reverse	GCTACGTGGTGCAAGAGTCA
rs11283943	1.5	55	Q-solution	Forward	TGGATGCACTTCTACCCTGA
				Reverse	CTAATCTGACAGGCCCACATC
rs11350445	1.5	55	Q-solution	Forward	CATTGTGTCCCTTTTCTCTCC
				Reverse	ACTGACTATTCAAGGGGATCG
rs11364237	1.5	55	Q-solution	Forward	ACCTCGTCCAATGATCCAAG
				Reverse	GACGATGTCCAACACAATGC
rs11376162	1.5	55	Q-solution	Forward	CTATTGTGGCCTCCCCAAG
				Reverse	CCCTTAGGATCTCCTTCCACA
rs11424286	1.5	55	Q-solution	Forward	GACACAGACATCAGGATGCAA
				Reverse	AGATCGCCAACAAGCTCTTC
rs11476163	1.5	55	Q-solution	Forward	ATCCCACCTACGGTCTCCTC
				Reverse	CTCAGTTTACCCTGCCTGCT
rs11477710	1.5	55	Q-solution	Forward	ACTTACCACTTGCACGCTCA
				Reverse	GCTGACATCATTTTTGTTTCTCC
rs11478027	1.5	55	Q-solution	Forward	AGGCCATTCTCCTCAGAAGC
				Reverse	GAAACGTGTTTCCTCACAAGC
rs11509437	2.5	55	Q-solution	Forward	GAGGGGGCCGATACAGTTAG
				Reverse	AAGTGACTTGGAAAGTGGGAAT
rs11564598	1.5	55	Q-solution	Forward	TGAGGCTGCAAGAAAGTTCA
				Reverse	TTAGGACAATGGCTGGCTTC
rs11564619	2.5	55	Q-solution	Forward	TTGTGAAGAAAACAGGCATTTG
				Reverse	GGGTCTTCTCACCGTGTTTG
rs140596	1.5	55	Q-solution	Forward	GCAGACATCGATGAGTGTGAA
				Reverse	TCTTCAGATGCCATGAATCC
rs17001464	1.5	55	Q-solution	Forward	GAGAACACAGGCGTGCATC
				Reverse	TCGGCTCTGTACCTGATGGT
rs17879749	1.5	55	Q-solution	Forward	TGCTCATGCTTTTCAACCAG
				Reverse	AGGGGTGGGAGAATTGAGAG
rs2066730	1.5	55	Q-solution	Forward	TTATTTTAGGTGTCTGGGGCAGT
				Reverse	GCCTCTGCTTTGTGAACCA
rs3069752	1.5	55	Q-solution	Forward	CGTTCTGGGCACTTACCTTT
				Reverse	CAGCAGCACAACCACCATAG

Locus	Mg^{2+}	Temp	Additive	Type	Primer
rs3092768	2.5	55	Q-solution	Forward	TGGCAGGAACTCTCTCTTCC
				Reverse	CAGACAAACCAAAGCAGAGGA
rs3212987	1.5	55	Q-solution	Forward	ATCTCCTCTTGGGGCACTGT
				Reverse	AGCCAGAGACAGAGGTGGTG
rs3214276	1.5	55	Q-solution	Forward	AACTCCCTCCAGAGCTACCA
				Reverse	TGGGTTTATGGCTCCATTGT
rs3217458	1.5	55	Q-solution	Forward	CTCTGCGGACCTTACCTCCT
				Reverse	ACTGTTCCATTCCCCACTTG
rs4067742	1.5	55	Q-solution	Forward	AGGCAGCCTGAAGTCCTACT
				Reverse	CATACCAGCAAAGGATGCAG
rs4987226	2.5	55	Q-solution	Forward	TCACTGGTAGAAAAGATTCTGTCTG
				Reverse	TTACCTGCATGGCTTCTCCT
rs5030651	1.5	55	Q-solution	Forward	GAGTACGGCCCTGAAGAAGA
				Reverse	GCGATTGCAGAAGATGACCT
rs5773188	1.5	55	Q-solution	Forward	AGGGGTGGGGAAAGAAGG
				Reverse	AGCATGCGGTCGAGAGAG
rs5790928	1.5	55	Q-solution	Forward	GGGATTGCACATGTTGTCTG
				Reverse	TGCAGAGGAAAGCTAAAGCA
rs5803440	1.5	55	Q-solution	Forward	TCCATCAGATGTTAGCACCAA
				Reverse	TGGGCAATGGGACAACTAAT
rs5814559	1.5	55	Q-solution	Forward	GGTGAAATCCAGTACCTCATCA
				Reverse	GCCCTTCACTTTCTGTTGCT
rs5820291	1.5	55	Q-solution	Forward	AACAGGCTTCATTGGAGAGG
				Reverse	GGGGTCTTCTGGATCTAGCC
rs5844947	1.5	55	Q-solution	Forward	GTCCACCCCAGGTAACACAG
				Reverse	AGCATGCAGCCAGAGGAG
rs5848002	2.5	55	Q-solution	Forward	AAAGACCTGCCGGGAAGA
				Reverse	CCAGGAGGGACTTACCAACA
rs5848302	1.5	55	Q-solution	Forward	AGAATGAGGAAACGGTGCTG
				Reverse	CTCTCCTCCCAACCCTGTTA
rs5883925	1.5	55	Q-solution	Forward	ACAAAGCGACCCAGCAAGT
				Reverse	AAGCATGTCAGCCACCTCTT
rs5888463	1.5	55	Q-solution	Forward	TCAGCCGGTAAGAGCTCACT
				Reverse	GGGTTTGGTCAGTAGCCACA
rs9332131	2.5	55	Q-solution	Forward	TGCTTCCTGATGAAAATGGA
				Reverse	CAACAAATCACAAATTCACAAGC
rs9332736	1.5	55	Q-solution	Forward	ACTGTTCGCAGAGTGTGTCG
				Reverse	TCAGCCTTTCCATCTTTGCT
rs9333357	1.5	55	Q-solution	Forward	CTCCTGCCTGAAGGACAGAC
				Reverse	TCTGCTGGATCATCTCATGG

References

1. Walther A, Johnstone E, Swanton C, Midgley R, Tomlinson I, Kerr D. Genetic prognostic and predictive markers in colorectal cancer. Nat Rev Cancer 2009;9:489-99.
2. Ferlay J, Bray F, Pisani P, Parkin DM. GLOBOCAN 2002: Cancer Incidence, Mortality and Prevalence Worldwide. IARC CancerBase No. 5. version 2.0 ed. Lyon: IARCPress; 2004.
3. Jemal A, Siegel R, Ward E, et al. Cancer statistics, 2008. CA Cancer J Clin 2008;58:71-96.
4. Williams PL, Warwick R, Dyson M, Bannister LH. Chapter 8: Splanchnology, Abdominal viscera. In: Gray's Anatomy. 37 ed. London: Churchill Livingstone; 1989.
5. Mak K. ICU web - Image library. http://wwwaiccuhkeduhk/web8/Hi%20res/Large%20boweljpg 2004.
6. CRUK. Key facts large bowel cancer (colorectal cancer). http://infocancerresearchukorg/cancerstats/types/bowel/ accessed 24 January 2009.
7. Whittemore AS, Wu-Williams AH, Lee M, et al. Diet, physical activity, and colorectal cancer among Chinese in North America and China. J Natl Cancer Inst 1990;82:915-26.
8. Armstrong B, Doll R. Environmental factors and cancer incidence and mortality in different countries, with special reference to dietary practices. Int J Cancer 1975;15:617-31.
9. Trock B, Lanza E, Greenwald P. Dietary fiber, vegetables, and colon cancer: critical review and meta-analyses of the epidemiologic evidence. J Natl Cancer Inst 1990;82:650-61.
10. Willett WC, Stampfer MJ, Colditz GA, Rosner BA, Speizer FE. Relation of meat, fat, and fiber intake to the risk of colon cancer in a prospective study among women. N Engl J Med 1990;323:1664-72.
11. Giovannucci E, Rimm EB, Stampfer MJ, Colditz GA, Ascherio A, Willett WC. Intake of fat, meat, and fiber in relation to risk of colon cancer in men. Cancer Res 1994;54:2390-7.
12. Slattery ML, Potter J, Caan B, et al. Energy balance and colon cancer--beyond physical activity. Cancer Res 1997;57:75-80.
13. Martinez ME, Giovannucci E, Spiegelman D, Hunter DJ, Willett WC, Colditz GA. Leisure-time physical activity, body size, and colon cancer in women. Nurses' Health Study Research Group. J Natl Cancer Inst 1997;89:948-55.
14. Giovannucci E, Rimm EB, Stampfer MJ, et al. A prospective study of cigarette smoking and risk of colorectal adenoma and colorectal cancer in U.S. men. J Natl Cancer Inst 1994;86:183-91.
15. Giovannucci E, Colditz GA, Stampfer MJ, et al. A prospective study of cigarette smoking and risk of colorectal adenoma and colorectal cancer in U.S. women. J Natl Cancer Inst 1994;86:192-9.
16. Lichtenstein P, Holm NV, Verkasalo PK, et al. Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. N Engl J Med 2000;343:78-85.
17. Merg A, Lynch HT, Lynch JF, Howe JR. Hereditary colon cancer--part I. Curr Probl Surg 2005;42:195-256.
18. Lockhart-Mummery JP. The Causation and Treatment of Multiple Adenomatosis of the Colon. Ann Surg 1934;99:178-84.
19. Bodmer WF, Bailey CJ, Bodmer J, et al. Localization of the gene for familial adenomatous polyposis on chromosome 5. Nature 1987;328:614-6.
20. Groden J, Thliveris A, Samowitz W, et al. Identification and characterization of the familial adenomatous polyposis coli gene. Cell 1991;66:589-600.
21. Kinzler KW, Nilbert MC, Su LK, et al. Identification of FAP locus genes from chromosome 5q21. Science 1991;253:661-5.

22. Beroud C, Soussi T. APC gene: database of germline and somatic mutations in human tumors and cell lines. Nucleic Acids Res 1996;24:121-4.
23. Bodmer W. Familial adenomatous polyposis (FAP) and its gene, APC. Cytogenet Cell Genet 1999;86:99-104.
24. Bisgaard ML, Fenger K, Bulow S, Niebuhr E, Mohr J. Familial adenomatous polyposis (FAP): frequency, penetrance, and mutation rate. Hum Mutat 1994;3:121-5.
25. Galiatsatos P, Foulkes WD. Familial adenomatous polyposis. Am J Gastroenterol 2006;101:385-98.
26. Sieber OM, Lipton L, Crabtree M, et al. Multiple colorectal adenomas, classic adenomatous polyposis, and germ-line mutations in MYH. N Engl J Med 2003;348:791-9.
27. Peltomaki P, de la Chapelle A. Mutations predisposing to hereditary nonpolyposis colorectal cancer. Adv Cancer Res 1997;71:93-119.
28. Grady WM, Carethers JM. Genomic and epigenetic instability in colorectal cancer pathogenesis. Gastroenterology 2008;135:1079-99.
29. Chung DC, Rustgi AK. The hereditary nonpolyposis colorectal cancer syndrome: genetics and clinical implications. Ann Intern Med 2003;138:560-70.
30. Vasen HF, Watson P, Mecklin JP, Lynch HT. New clinical criteria for hereditary nonpolyposis colorectal cancer (HNPCC, Lynch syndrome) proposed by the International Collaborative group on HNPCC. Gastroenterology 1999;116:1453-6.
31. Aaltonen LA, Salovaara R, Kristo P, et al. Incidence of hereditary nonpolyposis colorectal cancer and the feasibility of molecular screening for the disease. N Engl J Med 1998;338:1481-7.
32. Samowitz WS, Curtin K, Lin HH, et al. The colon cancer burden of genetically defined hereditary nonpolyposis colon cancer. Gastroenterology 2001;121:830-8.
33. Hemminki A, Markie D, Tomlinson I, et al. A serine/threonine kinase gene defective in Peutz-Jeghers syndrome. Nature 1998;391:184-7.
34. Olschwang S, Serova-Sinilnikova OM, Lenoir GM, Thomas G. PTEN germ-line mutations in juvenile polyposis coli. Nat Genet 1998;18:12-4.
35. Woodford-Richens K, Bevan S, Churchman M, et al. Analysis of genetic and phenotypic heterogeneity in juvenile polyposis. Gut 2000;46:656-60.
36. Howe JR, Bair JL, Sayed MG, et al. Germline mutations of the gene encoding bone morphogenetic protein receptor 1A in juvenile polyposis. Nat Genet 2001;28:184-7.
37. McCarthy MI, Abecasis GR, Cardon LR, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 2008;9:356-69.
38. Yeager M, Orr N, Hayes RB, et al. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. Nat Genet 2007;39:645-9.
39. Hunter DJ, Kraft P, Jacobs KB, et al. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nat Genet 2007;39:870-4.
40. Helgadottir A, Thorleifsson G, Manolescu A, et al. A common variant on chromosome 9p21 affects the risk of myocardial infarction. Science 2007;316:1491-3.
41. Zeggini E, Weedon MN, Lindgren CM, et al. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. Science 2007;316:1336-41.
42. Tomlinson IP, Webb E, Carvajal-Carmona L, et al. A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. Nat Genet 2008.
43. Jaeger E, Webb E, Howarth K, et al. Common genetic variants at the CRAC1 (HMPS) locus on chromosome 15q13.3 influence colorectal cancer risk. Nat Genet 2008;40:26-8.
44. Tomlinson I, Webb E, Carvajal-Carmona L, et al. A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. Nat Genet 2007;39:984-8.
45. Broderick P, Carvajal-Carmona L, Pittman AM, et al. A genome-wide association study shows that common alleles of SMAD7 influence colorectal cancer risk. Nat Genet 2007;39:1315-7.
46. Houlston RS, Webb E, Broderick P, et al. Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. Nat Genet 2008;40:1426-35.
47. McInerney N, Colleran G, Rowan A, et al. Low penetrance breast cancer predisposition SNPs are site specific. Breast Cancer Res Treat 2008.
48. Kiemeney LA, Thorlacius S, Sulem P, et al. Sequence variant on 8q24 confers susceptibility to urinary bladder cancer. Nat Genet 2008;40:1307-12.

49. Ghoussaini M, Song H, Koessler T, et al. Multiple loci with different cancer specificities within the 8q24 gene desert. J Natl Cancer Inst 2008;100:962-6.
50. Sansom OJ, Meniel VS, Muncan V, et al. Myc deletion rescues Apc deficiency in the small intestine. Nature 2007;446:676-9.
51. Nass SJ, Dickson RB. Defining a role for c-Myc in breast tumorigenesis. Breast Cancer Res Treat 1997;44:1-22.
52. Pittman AM, Webb E, Carvajal-Carmona L, et al. Refinement of the basis and impact of common 11q23.1 variation to the risk of developing colorectal cancer. Hum Mol Genet 2008;17:3720-7.
53. Kinzler KW, Vogelstein B. Lessons from hereditary colorectal cancer. Cell 1996;87:159-70.
54. Miyazaki M, Furuya T, Shiraki A, Sato T, Oga A, Sasaki K. The relationship of DNA ploidy to chromosomal instability in primary human colorectal cancers. Cancer Res 1999;59:5283-5.
55. Boland CR, Thibodeau SN, Hamilton SR, et al. 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 1998;58:5248-57.
56. Goel A, Arnold CN, Niedzwiecki D, et al. Characterization of sporadic colon cancer by patterns of genomic instability. Cancer Res 2003;63:1608-14.
57. Rowan A, Halford S, Gaasenbeek M, et al. Refining molecular analysis in the pathways of colorectal carcinogenesis. Clin Gastroenterol Hepatol 2005;3:1115-23.
58. Weisenberger DJ, Siegmund KD, Campan M, et al. CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer. Nat Genet 2006;38:787-93.
59. BRAF V599E becomes V600E. Catalogue Of Somatic Mutations In Cancer - (COSMIC) 2004.
60. Barault L, Charon-Barra C, Jooste V, et al. Hypermethylator phenotype in sporadic colon cancer: study on a population-based series of 582 cases. Cancer Res 2008;68:8541-6.
61. Ogino S, Nosho K, Kirkner GJ, et al. CpG island methylator phenotype, microsatellite instability, BRAF mutation and clinical outcome in colon cancer. Gut 2008.
62. Shen L, Toyota M, Kondo Y, et al. Integrated genetic and epigenetic analysis identifies three different subclasses of colon cancer. Proc Natl Acad Sci U S A 2007;104:18654-9.
63. Ogino S, Kawasaki T, Kirkner GJ, Loda M, Fuchs CS. CpG island methylator phenotype-low (CIMP-low) in colorectal cancer: possible associations with male sex and KRAS mutations. J Mol Diagn 2006;8:582-8.
64. Vogelstein B, Fearon ER, Hamilton SR, et al. Genetic alterations during colorectal-tumor development. N Engl J Med 1988;319:525-32.
65. Kinzler KW, Vogelstein B. Cancer-susceptibility genes. Gatekeepers and caretakers. Nature 1997;386:761, 3.
66. Powell SM, Zilz N, Beazer-Barclay Y, et al. APC mutations occur early during colorectal tumorigenesis. Nature 1992;359:235-7.
67. Kim JC, Koo KH, Roh SA, et al. Genetic and epigenetic changes in the APC gene in sporadic colorectal carcinoma with synchronous adenoma. Int J Colorectal Dis 2003;18:203-9.
68. Thorstensen L, Lind GE, Lovig T, et al. Genetic and epigenetic changes of components affecting the WNT pathway in colorectal carcinomas stratified by microsatellite instability. Neoplasia 2005;7:99-108.
69. Green RA, Kaplan KB. Chromosome instability in colorectal tumor cells is associated with defects in microtubule plus-end attachments caused by a dominant mutation in APC. J Cell Biol 2003;163:949-61.
70. Tighe A, Johnson VL, Albertella M, Taylor SS. Aneuploid colon cancer cells have a robust spindle checkpoint. EMBO Rep 2001;2:609-14.
71. Tighe A, Johnson VL, Taylor SS. Truncating APC mutations have dominant effects on proliferation, spindle checkpoint control, survival and chromosome stability. J Cell Sci 2004;117:6339-53.
72. Fodde R, Kuipers J, Rosenberg C, et al. Mutations in the APC tumour suppressor gene cause chromosomal instability. Nat Cell Biol 2001;3:433-8.
73. Kaplan KB, Burds AA, Swedlow JR, Bekir SS, Sorger PK, Nathke IS. A role for the Adenomatous Polyposis Coli protein in chromosome segregation. Nat Cell Biol 2001;3:429-32.
74. Shimizu Y, Ikeda S, Fujimori M, et al. Frequent alterations in the Wnt signaling pathway in colorectal cancer with microsatellite instability. Genes Chromosomes Cancer 2002;33:73-81.

75. Lovig T, Meling GI, Diep CB, et al. APC and CTNNB1 mutations in a large series of sporadic colorectal carcinomas stratified by the microsatellite instability status. Scand J Gastroenterol 2002;37:1184-93.
76. Luceri C, De Filippo C, Guglielmi F, et al. Microsatellite instability in a population of sporadic colorectal cancers: correlation between genetic and pathological profiles. Dig Liver Dis 2002;34:553-9.
77. Yarden Y, Sliwkowski MX. Untangling the ErbB signalling network. Nat Rev Mol Cell Biol 2001;2:127-37.
78. Doanes AM, Hegland DD, Sethi R, Kovesdi I, Bruder JT, Finkel T. VEGF stimulates MAPK through a pathway that is unique for receptor tyrosine kinases. Biochem Biophys Res Commun 1999;255:545-8.
79. Molloy CJ, Bottaro DP, Fleming TP, Marshall MS, Gibbs JB, Aaronson SA. PDGF induction of tyrosine phosphorylation of GTPase activating protein. Nature 1989;342:711-4.
80. Klint P, Kanda S, Claesson-Welsh L. Shc and a novel 89-kDa component couple to the Grb2-Sos complex in fibroblast growth factor-2-stimulated cells. J Biol Chem 1995;270:23337-44.
81. Bonni A, Brunet A, West AE, Datta SR, Takasu MA, Greenberg ME. Cell survival promoted by the Ras-MAPK signaling pathway by transcription-dependent and -independent mechanisms. Science 1999;286:1358-62.
82. Lavoie JN, L'Allemain G, Brunet A, Muller R, Pouyssegur J. Cyclin D1 expression is regulated positively by the p42/p44MAPK and negatively by the p38/HOGMAPK pathway. J Biol Chem 1996;271:20608-16.
83. Marshall CJ. Specificity of receptor tyrosine kinase signaling: transient versus sustained extracellular signal-regulated kinase activation. Cell 1995;80:179-85.
84. Andreyev HJ, Norman AR, Cunningham D, et al. Kirsten ras mutations in patients with colorectal cancer: the 'RASCAL II' study. Br J Cancer 2001;85:692-6.
85. Bos JL. ras oncogenes in human cancer: a review. Cancer Res 1989;49:4682-9.
86. Malkin D, Li FP, Strong LC, et al. Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. Science 1990;250:1233-8.
87. Hawkins N, Norrie M, Cheong K, et al. CpG island methylation in sporadic colorectal cancers and its relationship to microsatellite instability. Gastroenterology 2002;122:1376-87.
88. Balint EE, Vousden KH. Activation and activities of the p53 tumour suppressor protein. Br J Cancer 2001;85:1813-23.
89. Vousden KH, Lane DP. p53 in health and disease. Nat Rev Mol Cell Biol 2007;8:275-83.
90. Fearon ER, Cho KR, Nigro JM, et al. Identification of a chromosome 18q gene that is altered in colorectal cancers. Science 1990;247:49-56.
91. Tanaka K, Oshimura M, Kikuchi R, Seki M, Hayashi T, Miyaki M. Suppression of tumorigenicity in human colon carcinoma cells by introduction of normal chromosome 5 or 18. Nature 1991;349:340-2.
92. Alazzouzi H, Alhopuro P, Salovaara R, et al. SMAD4 as a prognostic marker in colorectal cancer. Clin Cancer Res 2005;11:2606-11.
93. Miyaki M, Iijima T, Konishi M, et al. Higher frequency of Smad4 gene mutation in human colorectal cancer with distant metastasis. Oncogene 1999;18:3098-103.
94. Riggins GJ, Kinzler KW, Vogelstein B, Thiagalingam S. Frequency of Smad gene mutations in human cancers. Cancer Res 1997;57:2578-80.
95. Salovaara R, Roth S, Loukola A, et al. Frequent loss of SMAD4/DPC4 protein in colorectal cancers. Gut 2002;51:56-9.
96. Takagi Y, Kohmura H, Futamura M, et al. Somatic alterations of the DPC4 gene in human colorectal cancers in vivo. Gastroenterology 1996;111:1369-72.
97. Derynck R, Akhurst RJ, Balmain A. TGF-beta signaling in tumor suppression and cancer progression. Nat Genet 2001;29:117-29.
98. Xu X, Brodie SG, Yang X, et al. Haploid loss of the tumor suppressor Smad4/Dpc4 initiates gastric polyposis and cancer in mice. Oncogene 2000;19:1868-74.
99. Grau AM, Datta PK, Zi J, Halder SK, Beauchamp RD. Role of Smad proteins in the regulation of NF-kappaB by TGF-beta in colon cancer cells. Cell Signal 2005.
100. Katuri V, Tang Y, Li C, et al. Critical interactions between TGF-beta signaling/ELF, and E-cadherin/beta-catenin mediated tumor suppression. Oncogene 2005.
101. Hannon GJ, Beach D. p15INK4B is a potential effector of TGF-beta-induced cell cycle arrest. Nature 1994;371:257-61.

102. Tang Y, Katuri V, Srinivasan R, et al. Transforming growth factor-beta suppresses nonmetastatic colon cancer through Smad4 and adaptor protein ELF at an early stage of tumorigenesis. Cancer Res 2005;65:4228-37.
103. Moustakas A, Souchelnytskyi S, Heldin CH. Smad regulation in TGF-beta signal transduction. J Cell Sci 2001;114:4359-69.
104. Oft M, Heider KH, Beug H. TGFbeta signaling is necessary for carcinoma cell invasiveness and metastasis. Curr Biol 1998;8:1243-52.
105. Portella G, Cumming SA, Liddell J, et al. Transforming growth factor beta is essential for spindle cell conversion of mouse skin carcinoma in vivo: implications for tumor invasion. Cell Growth Differ 1998;9:393-404.
106. Herman JG, Umar A, Polyak K, et al. Incidence and functional consequences of hMLH1 promoter hypermethylation in colorectal carcinoma. Proc Natl Acad Sci U S A 1998;95:6870-5.
107. Malkhosyan S, Rampino N, Yamamoto H, Perucho M. Frameshift mutator mutations. Nature 1996;382:499-500.
108. Sinicrope FA, Rego RL, Halling KC, et al. Prognostic impact of microsatellite instability and DNA ploidy in human colon carcinoma patients. Gastroenterology 2006;131:729-37.
109. Salahshor S, Kressner U, Pahlman L, Glimelius B, Lindmark G, Lindblom A. Colorectal cancer with and without microsatellite instability involves different genes. Genes Chromosomes Cancer 1999;26:247-52.
110. Samowitz WS, Holden JA, Curtin K, et al. Inverse relationship between microsatellite instability and K-ras and p53 gene alterations in colon cancer. Am J Pathol 2001;158:1517-24.
111. Ward R, Meagher A, Tomlinson I, et al. Microsatellite instability and the clinicopathological features of sporadic colorectal cancer. Gut 2001;48:821-9.
112. Li LS, Kim NG, Kim SH, et al. Chromosomal imbalances in the colorectal carcinomas with microsatellite instability. Am J Pathol 2003;163:1429-36.
113. Jass JR, Biden KG, Cummings MC, et al. Characterisation of a subtype of colorectal cancer combining features of the suppressor and mild mutator pathways. J Clin Pathol 1999;52:455-60.
114. Elsaleh H, Powell B, McCaul K, et al. P53 alteration and microsatellite instability have predictive value for survival benefit from chemotherapy in stage III colorectal carcinoma. Clin Cancer Res 2001;7:1343-9.
115. Samowitz WS, Albertsen H, Herrick J, et al. Evaluation of a large, population-based sample supports a CpG island methylator phenotype in colon cancer. Gastroenterology 2005;129:837-45.
116. Nagasaka T, Koi M, Kloor M, et al. Mutations in both KRAS and BRAF may contribute to the methylator phenotype in colon cancer. Gastroenterology 2008;134:1950-60, 60 e1.
117. Watanabe T, Wu TT, Catalano PJ, et al. Molecular predictors of survival after adjuvant chemotherapy for colon cancer. N Engl J Med 2001;344:1196-206.
118. Simms LA, Radford-Smith G, Biden KG, et al. Reciprocal relationship between the tumor suppressors p53 and BAX in primary colorectal cancers. Oncogene 1998;17:2003-8.
119. Ward RL, Cheong K, Ku SL, Meagher A, O'Connor T, Hawkins NJ. Adverse prognostic effect of methylation in colorectal cancer is reversed by microsatellite instability. J Clin Oncol 2003;21:3729-36.
120. Halford S, Sasieni P, Rowan A, et al. Low-level microsatellite instability occurs in most colorectal cancers and is a nonrandomly distributed quantitative trait. Cancer Res 2002;62:53-7.
121. Young J, Leggett B, Gustafson C, et al. Genomic instability occurs in colorectal carcinomas but not in adenomas. Hum Mutat 1993;2:351-4.
122. Segditsas S, Tomlinson I. Colorectal cancer and genetic alterations in the Wnt pathway. Oncogene 2006;25:7531-7.
123. Rajagopalan H, Bardelli A, Lengauer C, Kinzler KW, Vogelstein B, Velculescu VE. Tumorigenesis: RAF/RAS oncogenes and mismatch-repair status. Nature 2002;418:934.
124. Parsons R, Myeroff LL, Liu B, et al. Microsatellite instability and mutations of the transforming growth factor beta type II receptor gene in colorectal cancer. Cancer Res 1995;55:5548-50.
125. Schwartz S, Jr., Yamamoto H, Navarro M, Maestro M, Reventos J, Perucho M. Frameshift mutations at mononucleotide repeats in caspase-5 and other target genes in endometrial and gastrointestinal cancer of the microsatellite mutator phenotype. Cancer Res 1999;59:2995-3002.

126. Souza RF, Appel R, Yin J, et al. Microsatellite instability in the insulin-like growth factor II receptor gene in gastrointestinal tumours. Nat Genet 1996;14:255-7.
127. Markowitz S, Wang J, Myeroff L, et al. Inactivation of the type II TGF-beta receptor in colon cancer cells with microsatellite instability. Science 1995;268:1336-8.
128. Rampino N, Yamamoto H, Ionov Y, et al. Somatic frameshift mutations in the BAX gene in colon cancers of the microsatellite mutator phenotype. Science 1997;275:967-9.
129. Mao JH, Perez-Losada J, Wu D, et al. Fbxw7/Cdc4 is a p53-dependent, haploinsufficient tumour suppressor gene. Nature 2004;432:775-9.
130. Rajagopalan H, Jallepalli PV, Rago C, et al. Inactivation of hCDC4 can cause chromosomal instability. Nature 2004;428:77-81.
131. Kemp Z, Rowan A, Chambers W, et al. CDC4 mutations occur in a subset of colorectal cancers but are not predicted to cause loss of function and are not associated with chromosomal instability. Cancer Res 2005;65:11361-6.
132. Dukes C. The classification of cancer of the rectum. J Pathol Bacteriol 1932;35:323.
133. Astler VB, Coller FA. The prognostic significance of direct extension of carcinoma of the colon and rectum. Ann Surg 1954;139:846-52.
134. AJCC. AJCC Cancer Staging Manual. 6th ed. New York, NY: Springer; 2002.
135. O'Connell JB, Maggard MA, Ko CY. Colon cancer survival rates with the new American Joint Committee on Cancer sixth edition staging. J Natl Cancer Inst 2004;96:1420-5.
136. Libutti SK, L.B. S, A.K. R. Cancers of the Gastrointestinal Tract: Section 8: Cancer of the Colon. 7th ed. Philadelphia, PA: Lippincott, Williams and Wilkins; 2005.
137. Gray R, Barnwell J, McConkey C, Hills RK, Williams NS, Kerr DJ. Adjuvant chemotherapy versus observation in patients with colorectal cancer: a randomised study. Lancet 2007;370:2020-9.
138. Andre T, Boni C, Mounedji-Boudiaf L, et al. Oxaliplatin, fluorouracil, and leucovorin as adjuvant treatment for colon cancer. N Engl J Med 2004;350:2343-51.
139. Gill S, Loprinzi CL, Sargent DJ, et al. Pooled analysis of fluorouracil-based adjuvant therapy for stage II and III colon cancer: who benefits and by how much? J Clin Oncol 2004;22:1797-806.
140. Saltz LB, Niedzwiecki D, Hollis D, et al. Irinotecan fluorouracil plus leucovorin is not superior to fluorouracil plus leucovorin alone as adjuvant treatment for stage III colon cancer: results of CALGB 89803. J Clin Oncol 2007;25:3456-61.
141. Ychou M, Raoul J, Douillard J. A phase III randomized trial of LV5FU2+CPT-11 vs. LV5FU2 alone in adjuvant high risk colon cancer (FNCLCC Accord02/FFCD9802). J Clin Oncol (Meeting Abstracts) 2005;23:3502.
142. Van Cutsem E, Labianca R, Hossfeld G, et al. Randomized phase III trial comparing infused irinotecan/5-fluorouracil (5-FU)/folinic acid (IF) versus 5-FU/FA (F) in stage III colon cancer patients. Proc Am Soc Clin Oncol 2005;23:1090s.
143. Wolmark N, Yothers G, O'Connell M, et al. A phase III trial assessing bevacizumab in stage II and III carcinoma of the colon: Results of NSABP Protocol C-08. J Clin Oncol 2009;27:(suppl; abstr LBA4)
144. Combination Chemotherapy With or Without Bevacizumab in Treating Patients Who Have Undergone Surgery for Stage II or Stage III Colon Cancer. NCI Trial NCT00112918 2005.
145. Combination Chemotherapy With or Without Cetuximab in Treating Patients With Stage III Colon Cancer That Was Completely Removed By Surgery. NCI Trial NCT00265811 2005.
146. Compton CC, Fielding LP, Burgart LJ, et al. Prognostic factors in colorectal cancer. College of American Pathologists Consensus Statement 1999. Arch Pathol Lab Med 2000;124:979-94.
147. Muller S, Chesner IM, Egan MJ, et al. Significance of venous and lymphatic invasion in malignant polyps of the colon and rectum. Gut 1989;30:1385-91.
148. Greene FL, Stewart AK, Norton HJ. A new TNM staging strategy for node-positive (stage III) colon cancer: an analysis of 50,042 patients. Ann Surg 2002;236:416-21; discussion 21.
149. Steinberg SM, Barkin JS, Kaplan RS, Stablein DM. Prognostic indicators of colon tumors. The Gastrointestinal Tumor Study Group experience. Cancer 1986;57:1866-70.
150. Kemeny N, Braun DW, Jr. Prognostic factors in advanced colorectal carcinoma. Importance of lactic dehydrogenase level, performance status, and white blood cell count. Am J Med 1983;74:786-94.
151. Wanebo HJ, B. R, C.M. P. Preoperative carcinoembryonic antigenlevel as a prognostic indicator in colorectal cancer. . N Engl J Med 1978;299:448-51.

152. Giacchetti S, Itzhaki M, Gruia G, et al. Long-term survival of patients with unresectable colorectal cancer liver metastases following infusional chemotherapy with 5-fluorouracil, leucovorin, oxaliplatin and surgery. Ann Oncol 1999;10:663-9.
153. Cunningham D, Humblet Y, Siena S, et al. Cetuximab monotherapy and cetuximab plus irinotecan in irinotecan-refractory metastatic colorectal cancer. N Engl J Med 2004;351:337-45.
154. Hurwitz H, Fehrenbacher L, Novotny W, et al. Bevacizumab plus irinotecan, fluorouracil, and leucovorin for metastatic colorectal cancer. N Engl J Med 2004;350:2335-42.
155. Midgley R, Kerr DJ. Capecitabine: have we got the dose right? Nat Clin Pract Oncol 2009;6:17-24.
156. Midgley RS, Yanagisawa Y, Kerr DJ. Evolution of nonsurgical therapy for colorectal cancer. Nat Clin Pract Gastroenterol Hepatol 2009;6:108-20.
157. Locker GY, Hamilton S, Harris J, et al. ASCO 2006 update of recommendations for the use of tumor markers in gastrointestinal cancer. J Clin Oncol 2006;24:5313-27.
158. Fisher B, Redmond C, Fisher ER, Caplan R. Relative worth of estrogen or progesterone receptor and pathologic characteristics of differentiation as indicators of prognosis in node negative breast cancer patients: findings from National Surgical Adjuvant Breast and Bowel Project Protocol B-06. J Clin Oncol 1988;6:1076-87.
159. Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials. Lancet 2005;365:1687-717.
160. Vogel CL, Cobleigh MA, Tripathy D, et al. Efficacy and safety of trastuzumab as a single agent in first-line treatment of HER2-overexpressing metastatic breast cancer. J Clin Oncol 2002;20:719-26.
161. van 't Veer LJ, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002;415:530-6.
162. FDA. FDA Clears Breast Cancer Specific Molecular Prognostic Test. http://wwwfdagov/bbs/topics/NEWS/2007/NEW01555html 2007.
163. Andreyev HJ, Norman AR, Cunningham D, Oates JR, Clarke PA. Kirsten ras mutations in patients with colorectal cancer: the multicenter "RASCAL" study. J Natl Cancer Inst 1998;90:675-84.
164. Downward J. Targeting RAS signalling pathways in cancer therapy. Nat Rev Cancer 2003;3:11-22.
165. Stites EC, Trampont PC, Ma Z, Ravichandran KS. Network analysis of oncogenic Ras activation in cancer. Science 2007;318:463-7.
166. Eberhard DA, Johnson BE, Amler LC, et al. Mutations in the epidermal growth factor receptor and in KRAS are predictive and prognostic indicators in patients with non-small-cell lung cancer treated with chemotherapy alone and in combination with erlotinib. J Clin Oncol 2005;23:5900-9.
167. Karapetis CS, Khambata-Ford S, Jonker DJ, et al. K-ras mutations and benefit from cetuximab in advanced colorectal cancer. N Engl J Med 2008;359:1757-65.
168. Bokemeyer C, IBondarenko I, Hartmann JT, et al. KRAS status and efficacy of first-line treatment of patients with metastatic colorectal cancer (mCRC) with FOLFOX with or without cetuximab: The OPUS experience. J Clin Oncol 2008;26:abstr 4000.
169. Tejpar S, Peeters M, Humblet Y, et al. Relationship of efficacy with KRAS status (wild type versus mutant) in patients with irinotecan-refractory metastatic colorectal cancer (mCRC), treated with irinotecan (q2w) and escalating doses of cetuximab (q1w): The EVEREST experience. J Clin Oncol 2008;26:abstr 4001.
170. Van Cutsem E, Lang I, D'haens G. KRAS status and efficacy in the first-line treatment of patients with metastatic colorectal cancer (mCrc) treated with FOLFIRI with or without cetuximab: the CRYSTAL experience. J Clin Oncol 2008;26:abstr 2.
171. Van Cutsem E, Kohne CH, Hitre E, et al. Cetuximab and chemotherapy as initial treatment for metastatic colorectal cancer. N Engl J Med 2009;360:1408-17.
172. Amado RG, Wolf M, Peeters M, et al. Wild-type KRAS is required for panitumumab efficacy in patients with metastatic colorectal cancer. J Clin Oncol 2008;26:1626-34.
173. Di Nicolantonio F, Martini M, Molinari F, et al. Wild-Type BRAF Is Required for Response to Panitumumab or Cetuximab in Metastatic Colorectal Cancer. J Clin Oncol 2008.
174. Salomon DS, Brandt R, Ciardiello F, Normanno N. Epidermal growth factor-related peptides and their receptors in human malignancies. Crit Rev Oncol Hematol 1995;19:183-232.
175. Chung KY, Shia J, Kemeny NE, et al. Cetuximab shows activity in colorectal cancer patients with tumors that do not express the epidermal growth factor receptor by immunohistochemistry. J Clin Oncol 2005;23:1803-10.

176. Bondi J, Bukholm G, Nesland JM, Bukholm IR. Expression of non-membranous beta-catenin and gamma-catenin, c-Myc and cyclin D1 in relation to patient outcome in human colon adenocarcinomas. Apmis 2004;112:49-56.
177. Hugh TJ, Dillon SA, Taylor BA, Pignatelli M, Poston GJ, Kinsella AR. Cadherin-catenin expression in primary colorectal cancer: a survival analysis. Br J Cancer 1999;80:1046-51.
178. Russo A, Bazan V, Iacopetta B, Kerr D, Soussi T, Gebbia N. The TP53 colorectal cancer international collaborative study on the prognostic and predictive significance of p53 mutation: influence of tumor site, type of mutation, and adjuvant treatment. J Clin Oncol 2005;23:7518-28.
179. Munro AJ, Lain S, Lane DP. P53 abnormalities and outcomes in colorectal cancer: a systematic review. Br J Cancer 2005;92:434-44.
180. Popat S, Houlston RS. A systematic review and meta-analysis of the relationship between chromosome 18q genotype, DCC status and colorectal cancer prognosis. Eur J Cancer 2005;41:2060-70.
181. Halling KC, French AJ, McDonnell SK, et al. Microsatellite instability and 8p allelic imbalance in stage B2 and C colorectal cancers. J Natl Cancer Inst 1999;91:1295-303.
182. Roth AD, Tejpar S, Yan P, et al. Correlation of molecular markers in colon cancer with stage-specific prognosis: Results of the translational study on the PETACC 3 - EORTC 40993-SAKK 60-00 trial. ASCO Gastrointestinal Cancers Symposium 2009.
183. Carethers JM, Hawn MT, Greenson JK, Hitchcock CL, Boland CR. Prognostic significance of allelic lost at chromosome 18q21 for stage II colorectal cancer. Gastroenterology 1998;114:1188-95.
184. Alhopuro P, Alazzouzi H, Sammalkorpi H, et al. SMAD4 levels and response to 5-fluorouracil in colorectal cancer. Clin Cancer Res 2005;11:6311-6.
185. Boulay JL, Mild G, Lowy A, et al. SMAD4 is a predictive marker for 5-fluorouracil-based chemotherapy in patients with colorectal cancer. Br J Cancer 2002;87:630-4.
186. Olschwang S, Hamelin R, Laurent-Puig P, et al. Alternative genetic pathways in colorectal carcinogenesis. Proc Natl Acad Sci U S A 1997;94:12122-7.
187. Walther A, Houlston R, Tomlinson I. Association between chromosomal instability and prognosis in colorectal cancer: a meta-analysis. Gut 2008;57:941-50.
188. Popat S, Hubner R, Houlston RS. Systematic review of microsatellite instability and colorectal cancer prognosis. J Clin Oncol 2005;23:609-18.
189. Cheng YW, Pincas H, Bacolod MD, et al. CpG island methylator phenotype associates with low-degree chromosomal abnormalities in colorectal cancer. Clin Cancer Res 2008;14:6005-13.
190. Samowitz WS, Sweeney C, Herrick J, et al. Poor survival associated with the BRAF V600E mutation in microsatellite-stable colon cancers. Cancer Res 2005;65:6063-9.
191. Yamada Y, Jackson-Grusby L, Linhart H, et al. Opposing effects of DNA hypomethylation on intestinal and liver carcinogenesis. Proc Natl Acad Sci U S A 2005;102:13580-5.
192. Matsuzaki K, Deng G, Tanaka H, Kakar S, Miura S, Kim YS. The relationship between global methylation level, loss of heterozygosity, and microsatellite instability in sporadic colorectal cancer. Clin Cancer Res 2005;11:8564-9.
193. Rodriguez J, Frigola J, Vendrell E, et al. Chromosomal instability correlates with genome-wide DNA demethylation in human primary colorectal cancers. Cancer Res 2006;66:8462-9468.
194. Ribic CM, Sargent DJ, Moore MJ, et al. Tumor microsatellite-instability status as a predictor of benefit from fluorouracil-based adjuvant chemotherapy for colon cancer. N Engl J Med 2003;349:247-57.
195. Sargent DJ, Marsoni S, Thibodeau SN, et al. Confirmation of deficient mismatch repair (dMMR) as a predictive marker for lack of benefit from 5-FU based chemotherapy in stage II and III colon cancer (CC): A pooled molecular reanalysis of randomized chemotherapy trials. J Clin Oncol 2008; 26:abstr 4008.
196. Jo WS, Carethers JM. Chemotherapeutic implications in microsatellite unstable colorectal cancer. Cancer Biomark 2006;2:51-60.
197. Kim GP, Colangelo LH, Wieand HS, et al. Prognostic and predictive roles of high-degree microsatellite instability in colon cancer: a National Cancer Institute-National Surgical Adjuvant Breast and Bowel Project Collaborative Study. J Clin Oncol 2007;25:767-72.
198. Liang JT, Huang KC, Lai HS, et al. High-frequency microsatellite instability predicts better chemosensitivity to high-dose 5-fluorouracil plus leucovorin chemotherapy for stage IV sporadic colorectal cancer after palliative bowel resection. Int J Cancer 2002;101:519-25.

199. Bertagnolli MM, Niedzwiecki D, Compton CC, et al. Microsatellite instability predicts improved response to adjuvant therapy with irinotecan, fluorouracil, and leucovorin in stage III colon cancer: Cancer and Leukemia Group B Protocol 89803. J Clin Oncol 2009;27:1814-21.
200. Fallik D, Borrini F, Boige V, et al. Microsatellite instability is a predictive factor of the tumor response to irinotecan in patients with advanced colorectal cancer. Cancer Res 2003;63:5738-44.
201. Vilar E, Scaltriti M, Balmana J, et al. Microsatellite instability due to hMLH1 deficiency is associated with increased cytotoxicity to irinotecan in human colorectal cancer cell lines. Br J Cancer 2008;99:1607-12.
202. Swanton C, Marani M, Pardo O, et al. Regulators of mitotic arrest and ceramide metabolism are determinants of sensitivity to paclitaxel and other chemotherapeutic drugs. Cancer Cell 2007;11:498-512.
203. Sudo T, Nitta M, Saya H, Ueno NT. Dependence of paclitaxel sensitivity on a functional spindle assembly checkpoint. Cancer Res 2004;64:2502-8.
204. Parker WB, Cheng YC. Metabolism and mechanism of action of 5-fluorouracil. Pharmacol Ther 1990;48:381-95.
205. Popat S, Matakidou A, Houlston RS. Thymidylate synthase expression and prognosis in colorectal cancer: a systematic review and meta-analysis. J Clin Oncol 2004;22:529-36.
206. Marsh S, McKay JA, Cassidy J, McLeod HL. Polymorphism in the thymidylate synthase promoter enhancer region in colorectal cancer. Int J Oncol 2001;19:383-6.
207. Kawakami K, Salonga D, Park JM, et al. Different lengths of a polymorphic repeat sequence in the thymidylate synthase gene affect translational efficiency but not its gene expression. Clin Cancer Res 2001;7:4096-101.
208. Mandola MV, Stoehlmacher J, Muller-Weeks S, et al. A novel single nucleotide polymorphism within the 5' tandem repeat polymorphism of the thymidylate synthase gene abolishes USF-1 binding and alters transcriptional activity. Cancer Res 2003;63:2898-904.
209. Mandola MV, Stoehlmacher J, Zhang W, et al. A 6 bp polymorphism in the thymidylate synthase gene causes message instability and is associated with decreased intratumoral TS mRNA levels. Pharmacogenetics 2004;14:319-27.
210. Lenz HJ. Pharmacogenomics in colorectal cancer. Semin Oncol 2003;30:47-53.
211. Hitre E, Budai B, Adleff V, et al. Influence of thymidylate synthase gene polymorphisms on the survival of colorectal cancer patients receiving adjuvant 5-fluorouracil. Pharmacogenet Genomics 2005;15:723-30.
212. Lurje G, Zhang W, Yang D, et al. Thymidylate synthase haplotype is associated with tumor recurrence in stage II and stage III colon cancer. Pharmacogenet Genomics 2008;18:161-8.
213. Mattison LK, Fourie J, Desmond RA, Modak A, Saif MW, Diasio RB. Increased prevalence of dihydropyrimidine dehydrogenase deficiency in African-Americans compared with Caucasians. Clin Cancer Res 2006;12:5491-5.
214. Diasio RB, Johnson MR. The role of pharmacogenetics and pharmacogenomics in cancer chemotherapy with 5-fluorouracil. Pharmacology 2000;61:199-203.
215. Wei X, McLeod HL, McMurrough J, Gonzalez FJ, Fernandez-Salguero P. Molecular basis of the human dihydropyrimidine dehydrogenase deficiency and 5-fluorouracil toxicity. J Clin Invest 1996;98:610-5.
216. Etienne MC, Lagrange JL, Dassonville O, et al. Population study of dihydropyrimidine dehydrogenase in cancer patients. J Clin Oncol 1994;12:2248-53.
217. Ezzeldin HH, Diasio RB. Predicting fluorouracil toxicity: can we finally do it? J Clin Oncol 2008;26:2080-2.
218. Frosst P, Blom HJ, Milos R, et al. A candidate genetic risk factor for vascular disease: a common mutation in methylenetetrahydrofolate reductase. Nat Genet 1995;10:111-3.
219. Cohen V, Panet-Raymond V, Sabbaghian N, Morin I, Batist G, Rozen R. Methylenetetrahydrofolate reductase polymorphism in advanced colorectal cancer: a novel genomic predictor of clinical response to fluoropyrimidine-based chemotherapy. Clin Cancer Res 2003;9:1611-5.
220. Jakobsen A, Nielsen JN, Gyldenkerne N, Lindeberg J. Thymidylate synthase and methylenetetrahydrofolate reductase gene polymorphism in normal tissue as predictors of fluorouracil sensitivity. J Clin Oncol 2005;23:1365-9.
221. Sharma R, Hoskins JM, Rivory LP, et al. Thymidylate synthase and methylenetetrahydrofolate reductase gene polymorphisms and toxicity to capecitabine in advanced colorectal cancer patients. Clin Cancer Res 2008;14:817-25.

222. Etienne MC, Formento JL, Chazal M, et al. Methylenetetrahydrofolate reductase gene polymorphisms and response to fluorouracil-based treatment in advanced colorectal cancer patients. Pharmacogenetics 2004;14:785-92.
223. Etienne-Grimaldi MC, Francoual M, Formento JL, Milano G. Methylenetetrahydrofolate reductase (MTHFR) variants and fluorouracil-based treatments in colorectal cancer. Pharmacogenomics 2007;8:1561-6.
224. Mishima M, Samimi G, Kondo A, Lin X, Howell SB. The cellular pharmacology of oxaliplatin resistance. Eur J Cancer 2002;38:1405-12.
225. McIlwain CC, Townsend DM, Tew KD. Glutathione S-transferase polymorphisms: cancer incidence and therapy. Oncogene 2006;25:1639-48.
226. Lecomte T, Landi B, Beaune P, Laurent-Puig P, Loriot MA. Glutathione S-transferase P1 polymorphism (Ile105Val) predicts cumulative neuropathy in patients receiving oxaliplatin-based chemotherapy. Clin Cancer Res 2006;12:3050-6.
227. Kweekel DM, Gelderblom H, Guchelaar HJ. Pharmacology of oxaliplatin and the use of pharmacogenomics to individualize therapy. Cancer Treat Rev 2005;31:90-105.
228. Stoehlmacher J, Park DJ, Zhang W, et al. A multivariate analysis of genomic polymorphisms: prediction of clinical outcome to 5-FU/oxaliplatin combination chemotherapy in refractory colorectal cancer. Br J Cancer 2004;91:344-54.
229. McLeod HL, Sargent DJ, Marsh S, et al. Pharmacogenetic analysis of systemic toxicity and response after 5-fluorouracil (5FU)/CPT-11, 5FU/oxaliplatin (oxal), or CPT-11/oxal therapy for advanced colorectal cancer (CRC): Results from an intergroup trial Proc Am Soc Clin Oncol 2003;22:253, abstr 1013.
230. Iyer L, Hall D, Das S, et al. Phenotype-genotype correlation of in vitro SN-38 (active metabolite of irinotecan) and bilirubin glucuronidation in human liver tissue with UGT1A1 promoter polymorphism. Clinical pharmacology and therapeutics 1999;65:576-82.
231. Iyer L, Das S, Janisch L, et al. UGT1A1*28 polymorphism as a determinant of irinotecan disposition and toxicity. The pharmacogenomics journal 2002;2:43-7.
232. Hoskins JM, Goldberg RM, Qu P, Ibrahim JG, McLeod HL. UGT1A1*28 genotype and irinotecan-induced neutropenia: dose matters. J Natl Cancer Inst 2007;99:1290-5.
233. Liu CY, Chen PM, Chiou TJ, et al. UGT1A1*28 polymorphism predicts irinotecan-induced severe toxicities without affecting treatment outcome and survival in patients with metastatic colorectal carcinoma. Cancer 2008;112:1932-40.
234. Bandres E, Malumbres R, Cubedo E, et al. A gene signature of 8 genes could identify the risk of recurrence and progression in Dukes' B colon cancer patients. Oncol Rep 2007;17:1089-94.
235. Del Rio M, Molina F, Bascoul-Mollevi C, et al. Gene expression signature in advanced colorectal cancer patients select drugs and response for the use of leucovorin, fluorouracil, and irinotecan. J Clin Oncol 2007;25:773-80.
236. Grade M, Hormann P, Becker S, et al. Gene expression profiling reveals a massive, aneuploidy-dependent transcriptional deregulation and distinct differences between lymph node-negative and lymph node-positive colon carcinomas. Cancer Res 2007;67:41-56.
237. Kleivi K, Lind GE, Diep CB, et al. Gene expression profiles of primary colorectal carcinomas, liver metastases, and carcinomatoses. Mol Cancer 2007;6:2.
238. Barrier A, Boelle PY, Roser F, et al. Stage II colon cancer prognosis prediction by tumor gene expression profiling. J Clin Oncol 2006;24:4685-91.
239. Wang Y, Jatkoe T, Zhang Y, et al. Gene expression profiles and molecular markers to predict recurrence of Dukes' B colon cancer. J Clin Oncol 2004;22:1564-71.
240. Barrier A, Lemoine A, Boelle PY, et al. Colon cancer prognosis prediction by gene expression profiling. Oncogene 2005;24:6155-64.
241. Jiang Y, Casey G, Lavery IC, et al. Development of a clinically feasible molecular assay to predict recurrence of stage II colon cancer. J Mol Diagn 2008;10:346-54.
242. Lin YH, Friederichs J, Black MA, et al. Multiple gene expression classifiers from different array platforms predict poor prognosis of colorectal cancer. Clin Cancer Res 2007;13:498-507.
243. Barrier A, Roser F, Boelle PY, et al. Prognosis of stage II colon cancer by non-neoplastic mucosa gene expression profiling. Oncogene 2007;26:2642-8.

244. Barrier A, Boelle PY, Lemoine A, et al. Gene expression profiling of nonneoplastic mucosa may predict clinical outcome of colon cancer patients. Dis Colon Rectum 2005;48:2238-48.
245. Arango D, Laiho P, Kokko A, et al. Gene-expression profiling predicts recurrence in Dukes' C colorectal cancer. Gastroenterology 2005;129:874-84.
246. Frederiksen CM, Knudsen S, Laurberg S, Orntoft TF. Classification of Dukes' B and C colorectal cancers using expression arrays. J Cancer Res Clin Oncol 2003;129:263-71.
247. Schepeler T, Reinert JT, Ostenfeld MS, et al. Diagnostic and prognostic microRNAs in stage II colon cancer. Cancer Res 2008;68:6416-24.
248. Schetter AJ, Leung SY, Sohn JJ, et al. MicroRNA expression profiles associated with prognosis and therapeutic outcome in colon adenocarcinoma. Jama 2008;299:425-36.
249. Lanza G, Ferracin M, Gafa R, et al. mRNA/microRNA gene expression profile in microsatellite unstable colorectal cancer. Mol Cancer 2007;6:54.
250. Wang C, van Rijnsoever M, Grieu F, et al. Prognostic significance of microsatellite instability and Ki-ras mutation type in stage II colorectal cancer. Oncology 2003;64:259-65.
251. Pawitan Y, Michiels S, Koscielny S, Gusnanto A, Ploner A. False discovery rate, sensitivity and sample size for microarray studies. Bioinformatics 2005;21:3017-24.
252. Ntzani EE, Ioannidis JP. Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment. Lancet 2003;362:1439-44.
253. Michiels S, Koscielny S, Hill C. Prediction of cancer outcome with microarrays: a multiple random validation strategy. Lancet 2005;365:488-92.
254. Dupuy A, Simon RM. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. J Natl Cancer Inst 2007;99:147-57.
255. Ransohoff DF. Gene-expression signatures in breast cancer. N Engl J Med 2003;348:1715-7; author reply -7.
256. Simon R, Radmacher MD, Dobbin K, McShane LM. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. J Natl Cancer Inst 2003;95:14-8.
257. Ein-Dor L, Zuk O, Domany E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. Proc Natl Acad Sci U S A 2006;103:5923-8.
258. Lavery IC, Hammel J, Cowens J, et al. Relationship between tumor gene expression and recurrence in an observational cohort of patients with stage II/III colon cancer treated with surgery only: Quantitative RT-PCR assay of 375 genes in fixed paraffin-embedded (FPE) tissue. ASCO Gastrointestinal Cancer Symposium 2008:Abstr. 302.
259. Lusa L, McShane LM, Reid JF, et al. Challenges in projecting clustering results across gene expression-profiling datasets. J Natl Cancer Inst 2007;99:1715-23.
260. Lyman GH, Kuderer NM. Gene expression profile assays as predictors of recurrence-free survival in early-stage breast cancer: a metaanalysis. Clin Breast Cancer 2006;7:372-9.
261. Ioannidis JP. Microarrays and molecular research: noise discovery? Lancet 2005;365:454-5.
262. Ransohoff DF. Bias as a threat to the validity of cancer molecular-marker research. Nat Rev Cancer 2005;5:142-9.
263. He C, Kraft P, Chen C, et al. Genome-wide association studies identify loci associated with age at menarche and age at natural menopause. Nat Genet 2009;41:724-8.
264. Weedon MN, Lango H, Lindgren CM, et al. Genome-wide association analysis identifies 20 loci that influence adult height. Nat Genet 2008;40:575-83.
265. Hardy J, Singleton A. Genomewide association studies and human disease. N Engl J Med 2009;360:1759-68.
266. Carter SL, Eklund AC, Kohane IS, Harris LN, Szallasi Z. A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers. Nat Genet 2006;38:1043-8.
267. Peto R. Cancer, genes, and the environment. N Engl J Med 2000;343:1495; discussion -6.
268. Power C, Elliott J. Cohort profile: 1958 British birth cohort (National Child Development Study). Int J Epidemiol 2006;35:34-41.
269. Kosmider S, Jones I, Hibbert M, et al. Towards establishing a national colorectal cancer database: lessons learnt from Bio21 molecular medicine informatics model. ANZ J Surg 2008;78:803-9.

270. Steemers FJ, Chang W, Lee G, Barker DL, Shen R, Gunderson KL. Whole-genome genotyping with the single-base extension assay. Nat Methods 2006;3:31-3.
271. Illumina. Infinium II Assay Lab Setup and Procedures. Appendix A: System controls.
272. Illumina. Illumina GenCall Data Analysis Software. Illumina Tech Spotlight 2005.
273. Illumina. Infinium Genotyping Data Analysis. Illumina Technical Note 2007.
274. Skol AD, Scott LJ, Abecasis GR, Boehnke M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nat Genet 2006;38:209-13.
275. Devlin B, Roeder K. Genomic control for association studies. Biometrics 1999;55:997-1004.
276. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet 2006;2:e190.
277. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 2006;38:904-9.
278. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 2005;21:263-5.
279. Gabriel SB, Schaffner SF, Nguyen H, et al. The structure of haplotype blocks in the human genome. Science 2002;296:2225-9.
280. Lewontin RC. On measures of gametic disequilibrium. Genetics 1988;120:849-52.
281. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. Bmj 2003;327:557-60.
282. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. Bmj 1997;315:629-34.
283. Light R, Pillemer D. Summing Up: The science of Reviewing Research. Cambridge, Mass: Harvard University Press. 1984.
284. Thompson SG, Sharp SJ. Explaining heterogeneity in meta-analysis: a comparison of methods. Stat Med 1999;18:2693-708.
285. Marchini J, Cardon LR, Phillips MS, Donnelly P. The effects of human population structure on large genetic association studies. Nat Genet 2004;36:512-7.
286. Freedman ML, Reich D, Penney KL, et al. Assessing the impact of population stratification on genetic association studies. Nat Genet 2004;36:388-93.
287. Reich D, Price AL, Patterson N. Principal component analysis of genetic data. Nat Genet 2008;40:491-2.
288. Zhu XL, Liang L, Ding YQ. Overexpression of FMNL2 is closely related to metastasis of colorectal cancer. Int J Colorectal Dis 2008;23:1041-7.
289. Hayward DG, Fry AM. Nek2 kinase in chromosome instability and cancer. Cancer Lett 2006;237:155-66.
290. Scott RW, Olson MF. LIM kinases: function, regulation and association with human disease. J Mol Med 2007;85:555-68.
291. Wang L, Liu T, Wang Y, et al. Altered expression of desmocollin 3, desmoglein 3, and beta-catenin in oral squamous cell carcinoma: correlation with lymph node metastasis and cell proliferation. Virchows Arch 2007;451:959-66.
292. Wong MP, Cheang M, Yorida E, et al. Loss of desmoglein 1 expression associated with worse prognosis in head and neck squamous cell carcinoma patients. Pathology 2008;40:611-6.
293. Jung DJ, Sung HS, Goo YW, et al. Novel transcription coactivator complex containing activating signal cointegrator 1. Mol Cell Biol 2002;22:5203-11.
294. Sorensen PH, Lynch JC, Qualman SJ, et al. PAX3-FKHR and PAX7-FKHR gene fusions are prognostic indicators in alveolar rhabdomyosarcoma: a report from the children's oncology group. J Clin Oncol 2002;20:2672-9.
295. Jones AM, Mitter R, Poulsom R, et al. mRNA expression profiling of phyllodes tumours of the breast: identification of genes important in the development of borderline and malignant phyllodes tumours. J Pathol 2008;216:408-17.
296. Kim KA, Zhao J, Andarmani S, et al. R-Spondin proteins: a novel link to beta-catenin activation. Cell Cycle 2006;5:23-6.
297. Farrell C, Crimm H, Meeh P, et al. Somatic mutations to CSMD1 in colorectal adenocarcinomas. Cancer Biol Ther 2008;7:609-13.

298. Scholnick SB, Richter TM. The role of CSMD1 in head and neck carcinogenesis. Genes Chromosomes Cancer 2003;38:281-3.
299. Zanke BW, Greenwood CM, Rangrej J, et al. Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. Nat Genet 2007;39:989-94.
300. Bresalier RS, Sandler RS, Quan H, et al. Cardiovascular events associated with rofecoxib in a colorectal adenoma chemoprevention trial. N Engl J Med 2005;352:1092-102.
301. Zaykin DV, Meng Z, Ehm MG. Contrasting linkage-disequilibrium patterns between cases and controls as a novel association-mapping method. Am J Hum Genet 2006;78:737-46.
302. Gillen DL, Emerson SS. Information Growth in a Family of Weighted Logrank Statistics Under Repeated Analyses. Sequential Analysis 2005;24:1-22.
303. Cuzick J. A method for analysing case-control studies with ordinal exposure variables. Biometrics 1985;41:609-21.
304. Gallo KA, Johnson GL. Mixed-lineage kinase control of JNK and p38 MAPK pathways. Nat Rev Mol Cell Biol 2002;3:663-72.
305. Dhanasekaran DN, Johnson GL. MAPKs: function, regulation, role in cancer and therapeutic targeting. Oncogene 2007;26:3097-9.
306. Fang JY, Richardson BC. The MAPK signalling pathways and colorectal cancer. The Lancet Oncology 2005;6:322-7.
307. Volkmann N, Amann KJ, Stoilova-McPhie S, et al. Structure of Arp2/3 complex in its activated state and in actin filament branch junctions. Science 2001;293:2456-9.
308. Vavouri T, McEwen GK, Woolfe A, Gilks WR, Elgar G. Defining a genomic radius for long-range enhancer action: duplicated conserved non-coding elements hold the key. Trends Genet 2006;22:5-10.
309. Faraway JJ. On the cost of data analysis. J Comput Graph Stat 1992;1:213-29.
310. Goring HH, Terwilliger JD, Blangero J. Large upward bias in estimation of locus-specific effects from genomewide scans. Am J Hum Genet 2001;69:1357-69.
311. Davis CE. The effect of regression to the mean in epidemiologic and clinical studies. Am J Epidemiol 1976;104:493-8.
312. Yamaguchi H, Condeelis J. Regulation of the actin cytoskeleton in cancer cell migration and invasion. Biochimica et Biophysica Acta (BBA) - Molecular Cell Research 2007;1773:642-52.
313. Haiman CA, Le Marchand L, Yamamato J, et al. A common genetic risk factor for colorectal and prostate cancer. Nat Genet 2007;39:954-6.
314. Diversity and heterogeneity. The Cochrane Collaboration open learning material 2002.
315. Sargent DJ, Wieand HS, Haller DG, et al. Disease-free survival versus overall survival as a primary end point for adjuvant colon cancer studies: individual patient data from 20,898 patients on 18 randomized trials. J Clin Oncol 2005;23:8664-70.
316. Liu P, Pazin DE, Merson RR, Albrecht KH, Vaziri C. The developmentally-regulated Smoc2 gene is repressed by Aryl-hydrocarbon receptor (Ahr) signaling. Gene 2009;433:72-80.
317. Kopnin PB, Agapova LS, Kopnin BP, Chumakov PM. Repression of sestrin family genes contributes to oncogenic Ras-induced reactive oxygen species up-regulation and genetic instability. Cancer Res 2007;67:4671-8.
318. Zighelboim I, Goodfellow PJ, Schmidt AP, et al. Differential methylation hybridization array of endometrial cancers reveals two novel cancer-specific methylation markers. Clin Cancer Res 2007;13:2882-9.
319. Eswaran J, von Kries JP, Marsden B, et al. Crystal structures and inhibitor identification for PTPN5, PTPRR and PTPN7: a family of human MAPK-specific protein tyrosine phosphatases. Biochem J 2006;395:483-91.
320. Yamamoto H, Kishimoto T, Minamoto S. NF-kappaB activation in CD27 signaling: involvement of TNF receptor-associated factors in its signaling and identification of functional region of CD27. J Immunol 1998;161:4753-9.
321. Gabrilovich DI, Cheng P, Fan Y, et al. H1(0) histone and differentiation of dendritic cells. A molecular target for tumor-derived factors. J Leukoc Biol 2002;72:285-96.
322. Liu L, Channavajhala PL, Rao VR, et al. Proteomic characterization of the dynamic KSR-2 interactome, a signaling scaffold complex in MAPK pathway. Biochim Biophys Acta 2009.

323. Razidlo GL, Johnson HJ, Stoeger SM, Cowan KH, Bessho T, Lewis RE. KSR1 is required for cell cycle reinitiation following DNA damage. J Biol Chem 2009;284:6705-15.
324. Maccarana M, Olander B, Malmstrom J, et al. Biosynthesis of dermatan sulfate: chondroitin-glucuronate C5-epimerase is identical to SART2. J Biol Chem 2006;281:11560-8.
325. Sasatomi T, Suefuji Y, Matsunaga K, et al. Expression of tumor rejection antigens in colorectal carcinomas. Cancer 2002;94:1636-41.
326. Haiman CA, Patterson N, Freedman ML, et al. Multiple regions within 8q24 independently affect risk for prostate cancer. Nat Genet 2007;39:638-44.
327. Smith C, Butler JA. Colorectal cancer in patients younger than 40 years of age. Dis Colon Rectum 1989;32:843-6.
328. D'Onofrio GM, Tan EG. Is colorectal carcinoma in the young a more deadly disease? Aust N Z J Surg 1985;55:537-40.
329. Quah HM, Joseph R, Schrag D, et al. Young age influences treatment but not outcome of colon cancer. Ann Surg Oncol 2007;14:2759-65.
330. Mitry E, Benhamiche AM, Jouve JL, Clinard F, Finn-Faivre C, Faivre J. Colorectal adenocarcinoma in patients under 45 years of age: comparison with older patients in a well-defined French population. Dis Colon Rectum 2001;44:380-7.
331. Anders CK, Hsu DS, Broadwater G, et al. Young age at diagnosis correlates with worse prognosis and defines a subset of breast cancers with shared patterns of gene expression. J Clin Oncol 2008;26:3324-30.
332. Pomerantz MM, Ahmadiyeh N, Jia L, et al. The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. Nat Genet 2009;41:882-4.
333. Prochownik EV. c-Myc: linking transformation and genomic instability. Curr Mol Med 2008;8:446-58.
334. Prochownik EV, Li Y. The ever expanding role for c-Myc in promoting genomic instability. Cell Cycle 2007;6:1024-9.
335. Inghirami G, Grignani F, Sternas L, Lombardi L, Knowles DM, Dalla-Favera R. Down-regulation of LFA-1 adhesion receptors by C-myc oncogene in human B lymphoblastoid cells. Science 1990;250:682-6.
336. He TC, Sparks AB, Rago C, et al. Identification of c-MYC as a target of the APC pathway. Science 1998;281:1509-12.
337. Tuupanen S, Turunen M, Lehtonen R, et al. The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. Nat Genet 2009;41:885-90.
338. Roussel MF. Key effectors of signal transduction and G1 progression. Adv Cancer Res 1998;74:1-24.
339. Boxer LM, Dang CV. Translocations involving c-myc and c-myc function. Oncogene 2001;20:5595-610.
340. Juin P, Hueber AO, Littlewood T, Evan G. c-Myc-induced sensitization to apoptosis is mediated through cytochrome c release. Genes Dev 1999;13:1367-81.
341. Barr LF, Campbell SE, Bochner BS, Dang CV. Association of the decreased expression of alpha3beta1 integrin with the altered cell: environmental interactions and enhanced soft agar cloning ability of c-myc-overexpressing small cell lung cancer cells. Cancer Res 1998;58:5537-45.
342. Huh YO, Lin KI, Vega F, et al. MYC translocation in chronic lymphocytic leukaemia is associated with increased prolymphocytes and a poor prognosis. Br J Haematol 2008;142:36-44.
343. Mossafa H, Damotte D, Jenabian A, et al. Non-Hodgkin's lymphomas with Burkitt-like cells are associated with c-Myc amplification and poor prognosis. Leuk Lymphoma 2006;47:1885-93.
344. Kanungo A, Medeiros LJ, Abruzzo LV, Lin P. Lymphoid neoplasms associated with concurrent t(14;18) and 8q24/c-MYC translocation generally have a poor prognosis. Mod Pathol 2006;19:25-33.
345. Cooper GM, Zerr T, Kidd JM, Eichler EE, Nickerson DA. Systematic assessment of copy number variant detection via genome-wide SNP genotyping. Nat Genet 2008;40:1199-203.
346. Affymetrix. GeneChip® Human Mapping 500K Array Set. Affymetrix product family, Data sheet 2006.
347. Illumina. The Power of Intelligent SNP Selection. Illumina Technical Note 2008.
348. Frazer KA, Ballinger DG, Cox DR, et al. A second generation human haplotype map of over 3.1 million SNPs. Nature 2007;449:851-61.
349. Samowitz WS, Curtin K, Leppert MF, Slattery ML. Uncommon TGFBRI allele is not associated with increased susceptibility to colon cancer. Genes Chromosomes Cancer 2001;32:381-3.

350. Daley D, Morgan W, Lewis S, et al. Is TGFBR1*6A a susceptibility allele for nonsyndromic familial colorectal neoplasia? Cancer Epidemiol Biomarkers Prev 2007;16:892-4.
351. Stefanovska AM, Efremov GD, Dimovski AJ, et al. TbetaR-I(6A) polymorphism is not a tumor susceptibility allele in Macedonian colorectal cancer patients. Correspondence re: B. Pasche et al. Type I TbetaR-I(6A) Is a Candidate Tumor Susceptibility Allele. Cancer Res., 58: 2727-2732, 1998. Cancer Res 2001;61:8351-2.
352. Kaklamani VG, Hou N, Bian Y, et al. TGFBR1*6A and cancer risk: a meta-analysis of seven case-control studies. J Clin Oncol 2003;21:3236-43.
353. Lai R. Association between TGFBR1*6A and cancer: is there any evidence? J Clin Oncol 2004;22:2754; author reply -5.
354. Pasche B, Kaklamani V, Hou N, et al. TGFBR1*6A and cancer: a meta-analysis of 12 case-control studies. J Clin Oncol 2004;22:756-8.
355. Skoglund J, Song B, Dalen J, et al. Lack of an association between the TGFBR1*6A variant and colorectal cancer risk. Clin Cancer Res 2007;13:3748-52.
356. White S, Bubb VJ, Wyllie AH. Germline APC mutation (Gln1317) in a cancer-prone family that does not result in familial adenomatous polyposis. Genes Chromosomes Cancer 1996;15:122-8.
357. Frayling IM, Beck NE, Ilyas M, et al. 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 1998;95:10722-7.
358. Popat S, Stone J, Coleman G, et al. Prevalence of the APC E1317Q variant in colorectal cancer patients. Cancer Lett 2000;149:203-6.
359. Koch A, Denkhaus D, Albrecht S, Leuschner I, von Schweinitz D, Pietsch T. Childhood hepatoblastomas frequently carry a mutated degradation targeting box of the beta-catenin gene. Cancer Res 1999;59:269-73.
360. Zeggini E, Rayner W, Morris AP, et al. An evaluation of HapMap sample size and tagging SNP performance in large-scale empirical and simulated data sets. Nat Genet 2005;37:1320-2.
361. International Consortium Announces the 1000 Genomes Project. http://www1000genomesorg/pagephp?page=home 2008.
362. Song H, Ramus SJ, Tyrer J, et al. A genome-wide association study identifies a new ovarian cancer susceptibility locus on 9p22.2. Nat Genet 2009;advance online publication.
363. Stacey SN, Sulem P, Masson G, et al. New common variants affecting susceptibility to basal cell carcinoma. Nat Genet 2009;41:909-14.
364. Skibola CF, Bracci PM, Halperin E, et al. Genetic variants at 6p21.33 are associated with susceptibility to follicular lymphoma. Nat Genet 2009;41:873-5.
365. Papaemmanuil E, Hosking FJ, Vijayakrishnan J, et al. Loci on 7p12.2, 10q21.2 and 14q11.2 are associated with risk of childhood acute lymphoblastic leukemia. Nat Genet 2009;advance online publication.
366. Bishop DT, Demenais F, Iles MM, et al. Genome-wide association study identifies three loci associated with melanoma risk. Nat Genet 2009;41:920-5.
367. Shete S, Hosking FJ, Robertson LB, et al. Genome-wide association study identifies five susceptibility loci for glioma. Nat Genet 2009;41:899-904.
368. Newton-Cheh C, Johnson T, Gateva V, et al. Genome-wide association study identifies eight loci associated with blood pressure. Nat Genet 2009;41:666-76.
369. Barrett JC, Clayton DG, Concannon P, et al. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nat Genet 2009;41:703-7.
370. De Jager PL, Jia X, Wang J, et al. Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci. Nat Genet 2009;41:776-82.
371. Sasieni PD. From genotypes to genes: doubling the sample size. Biometrics 1997;53:1253-61.
372. Tenesa A, Farrington SM, Prendergast JG, et al. Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. Nat Genet 2008;40:631-7.
373. Otsubo T, Iwaya K, Mukai Y, et al. Involvement of Arp2//3 complex in the process of colorectal carcinogenesis. Mod Pathol 2004;17:461-7.
374. Bundschu K, Walter U, Schuh K. Getting a first clue about SPRED functions. BioEssays 2007;29:897-907.

375. Kotzsch M, Sieuwerts AM, Grosser M, et al. Urokinase receptor splice variant uPAR-del4/5-associated gene expression in breast cancer: identification of rab31 as an independent prognostic factor. Breast Cancer Res Treat 2008;111:229-40.
376. Potapova IA, Cohen IS, Doronin SV. Voltage-gated ion channel Kv4.3 is associated with Rap guanine nucleotide exchange factors and regulates angiotensin receptor type 1 signaling to small G-protein Rap. FEBS J 2007;274:4375-84.
377. Goldstein DB. Common Genetic Variation and Human Traits. N Engl J Med 2009;360:1696-8.
378. McCarroll SA. Extending genome-wide association studies to copy-number variation. Hum Mol Genet 2008;17:R135-42.
379. Iafrate AJ, Feuk L, Rivera MN, et al. Detection of large-scale variation in the human genome. Nat Genet 2004;36:949-51.
380. McCarroll SA, Hadnott TN, Perry GH, et al. Common deletion polymorphisms in the human genome. Nat Genet 2006;38:86-92.
381. Redon R, Ishikawa S, Fitch KR, et al. Global variation in copy number in the human genome. Nature 2006;444:444-54.
382. Sebat J, Lakshmi B, Troge J, et al. Large-Scale Copy Number Polymorphism in the Human Genome. Science 2004;305:525-8.
383. Tuzun E, Sharp AJ, Bailey JA, et al. Fine-scale structural variation of the human genome. Nat Genet 2005;37:727-32.
384. McCarroll SA, Kuruvilla FG, Korn JM, et al. Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat Genet 2008;40:1166-74.
385. Greshock J, Feng B, Nogueira C, et al. A Comparison of DNA Copy Number Profiling Platforms. Cancer Res 2007;67:10173-80.
386. Lin M, Wei L-J, Sellers WR, Lieberfarb M, Wong WH, Li C. dChipSNP: significance curve and clustering of SNP-array-based loss-of-heterozygosity data. Bioinformatics 2004;20:1233-40.
387. Liu W, Sun J, Li G, et al. Association of a Germ-Line Copy Number Variation at 2p24.3 and Risk for Aggressive Prostate Cancer. Cancer Res 2009;69:2176-9.
388. Bienz M, Clevers H. Linking colorectal cancer to Wnt signaling. Cell 2000;103:311-20.
389. Folsom AR, Pankow JS, Peacock JM, Bielinski SJ, Heiss G, Boerwinkle E. Variation in TCF7L2 and increased risk of colon cancer: the Atherosclerosis Risk in Communities (ARIC) Study. Diabetes Care 2008;31:905-9.
390. Hazra A, Fuchs CS, Chan AT, Giovannucci EL, Hunter DJ. Association of the TCF7L2 polymorphism with colorectal cancer and adenoma risk. Cancer Causes Control 2008;19:975-80.
391. Yeager M, Xiao N, Hayes RB, et al. Comprehensive resequence analysis of a 136 kb region of human chromosome 8q24 associated with prostate and colon cancers. Hum Genet 2008;124:161-70.
392. Conrad DF, Andrews TD, Carter NP, Hurles ME, Pritchard JK. A high-resolution survey of deletion polymorphism in the human genome. Nat Genet 2006;38:75-81.
393. McCarthy MI, Hirschhorn JN. Genome-wide association studies: potential next steps on a genetic journey. Hum Mol Genet 2008;17:R156-65.
394. Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. Nat Genet 2008;40:695-701.
395. Kerkel K, Spadola A, Yuan E, et al. Genomic surveys by methylation-sensitive SNP analysis identify sequence-dependent allele-specific DNA methylation. Nat Genet 2008;40:904-8.
396. Hitchins MP, Wong JJ, Suthers G, et al. Inheritance of a cancer-associated MLH1 germ-line epimutation. N Engl J Med 2007;356:697-705.
397. Schuster SC. Next-generation sequencing transforms today's biology. Nat Methods 2008;5:16-8.
398. Hoggart CJ, Clark TG, De Iorio M, Whittaker JC, Balding DJ. Genome-wide significance for dense SNP and resequencing data. Genet Epidemiol 2008;32:179-85.
399. Kanetsky PA, Mitra N, Vardhanabhuti S, et al. Common variation in KITLG and at 5q31.3 predisposes to testicular germ cell cancer. Nat Genet 2009;41:811-5.
400. Park H, Kim Y, Lim Y, Han I, Oh ES. Syndecan-2 mediates adhesion and proliferation of colon carcinoma cells. J Biol Chem 2002;277:29730-6.
401. Hu G, Chong RA, Yang Q, et al. MTDH activation by 8q22 genomic gain promotes chemoresistance and metastasis of poor-prognosis breast cancer. Cancer Cell 2009;15:9-20.

402. O'Brien PM, Davies MJ, Scurry JP, et al. The E3 ubiquitin ligase EDD is an adverse prognostic factor for serous epithelial ovarian cancer and modulates cisplatin resistance in vitro. Br J Cancer 2008;98:1085-93.
403. Amson RB, Nemani M, Roperch JP, et al. Isolation of 10 differentially expressed cDNAs in p53-induced apoptosis: activation of the vertebrate homologue of the drosophila seven in absentia gene. Proc Natl Acad Sci U S A 1996;93:3953-7.
404. Passer BJ, Nancy-Portebois V, Amzallag N, et al. The p53-inducible TSAP6 gene product regulates apoptosis and the cell cycle and interacts with Nix and the Myt1 kinase. Proc Natl Acad Sci U S A 2003;100:2284-9.
405. Wang Z, Shen D, Parsons DW, et al. Mutational analysis of the tyrosine phosphatome in colorectal cancers. Science 2004;304:1164-6.
406. Castro CP, Piscopo D, Nakagawa T, Derynck R. Cornichon regulates transport and secretion of TGFalpha-related proteins in metazoan cells. J Cell Sci 2007;120:2454-66.
407. Italiano A, Ebran N, Attias R, et al. NFIB rearrangement in superficial, retroperitoneal, and colonic lipomas with aberrations involving chromosome band 9p22. Genes Chromosomes Cancer 2008;47:971-7.
408. Kalfa TA, Thull JD, Butkowski RJ, Charonis AS. Tubulointerstitial nephritis antigen interacts with laminin and type IV collagen and promotes cell adhesion. J Biol Chem 1994;269:1654-9.
409. Scarlett DJ, Herst PM, Berridge MV. Multiple proteins with single activities or a single protein with multiple activities: the conundrum of cell surface NADH oxidoreductases. Biochim Biophys Acta 2005;1708:108-19.
410. Medina MA, del Castillo-Olivares A, Nunez de Castro I. Multifunctional plasma membrane redox systems. BioEssays 1997;19:977-84.
411. Beak JY, Kang HS, Kim YS, Jetten AM. Kruppel-like zinc finger protein Glis3 promotes osteoblast differentiation by regulating FGF18 expression. J Bone Miner Res 2007;22:1234-44.
412. Craven RJ, Cance WG, Liu ET. The nuclear tyrosine kinase Rak associates with the retinoblastoma protein pRb. Cancer Res 1995;55:3969-72.
413. Meyer T, Xu L, Chang J, Liu ET, Craven RJ, Cance WG. Breast cancer cell line proliferation blocked by the Src-related Rak tyrosine kinase. International Journal of Cancer 2003;104:139-46.
414. Hennemann H, Vassen L, Geisen C, Eilers M, Moroy T. Identification of a novel Kruppel-associated box domain protein, Krim-1, that interacts with c-Myc and inhibits its oncogenic activity. J Biol Chem 2003;278:28799-811.
415. Habuchi H, Tanaka M, Habuchi O, et al. The occurrence of three isoforms of heparan sulfate 6-O-sulfotransferase having different specificities for hexuronic acid adjacent to the targeted N-sulfoglucosamine. J Biol Chem 2000;275:2859-68.