University of Oxford

Germline Determinants of Colorectal Cancer Risk and Outcome

Luke Freeman-Mills

Merton College

A thesis submitted for the degree of

Doctor of Philosophy

Hilary Term 2017

Germline Determinants of Colorectal Cancer Risk and Outcome

Luke Freeman-Mills
Merton College, University of Oxford, Hilary 2017

Abstract

Colorectal cancer (CRC) is the third most common global cancer. Approximately one fifth of phenotypic variance in CRC is attributable to additive genetic inputs: the disease risk is heritable. The nature and location of these genetic inputs has been the preoccupation of association studies over the last decade. Variation at some 30 genomic loci influences sporadic CRC risk. However, in Mendelian disorders this risk is almost absolute. Such cancer syndromes include Lynch syndrome, caused by defective DNA mismatch repair, and serrated polyposis syndrome, which has no confirmed genetic cause.

The aim of this thesis was to continue exploring how common and rare germline genetics influence CRC liability. A reexamination of the chromosome 15 *GREM1* locus in seven imputed UK case-control cohorts revealed a new, genome-wide significant risk polymorphism at rs17816465, independent of previously reported associations. This polymorphism lies near to conserved regulatory elements within the *FMN1* gene and had allele-specific luciferase activity.

Similar fine-mapping at the *POLD3* locus showed a large haplotype carrying CRC risk, making further refinement of the signal difficult. A risk score incorporating a correlate of this haplotype and all other independently CRC-associated polymorphisms was significantly higher in cases than in controls, but did not predict patient survival. Instead, the risk score anticorrelated with tumour driver mutation burden in publically available colorectal cancer data.

This thesis also reports the mapping and calling of the whole-genome sequences of 299 UK CRC and adenoma patients. Rare variants within this cohort were shown to be shared beyond expectation in geographic clines not visible to principal component analysis. Future rare variant association studies will need to correct for cryptic population structure using other statistical tools.

In contravention of recruitment criteria, known polyposis syndrome genes were often deleteriously mutant within the genomes. Also, filtering on phenotype or gene subsets revealed potentially pathogenic variants in either whole-genome or whole-exome sequences in *EXO1*, *RPA4*, and *LNX1*. One individual displayed homozygous mutation of the mismatch repair gene *MSH3*. This was the second account of a germline *MSH3* change causing tetranucleotide-unstable CRC. Lastly, a family suffering serrated polyposis was shown to carry a heterozygous *RNF43* frameshift that segregated with the disease.

This thesis improves our account of both sporadic and Mendelian CRC risk. It identifies a new genome-wide significant association with sporadic CRC at the *GREM1* locus and provides a haplotypic account of risk at *POLD3*. It also describes two rare germline causes of Mendelian CRC: *MSH3* and *RNF43* loss of function.

Acknowledgements

I feel astonishingly lucky to have spent four years in as friendly, kind, and stimulating a group as the Tomlinson lab. In particular, Len, Tim, Xiaowei, David, Enric, Shazia, Hayley, Emma, Danny, Emily, American Dan, Sarah, Chaido, Rachael, Roland, and Chiara have provided friendship and support. Thank you all.

Friends near and far have been wonderful. For your patience and your companionship: Alex, Dennis, Beau, Hugo, Katie, Helen, Johannes, Daniel, Dan, Joe, Jessica, and Arthur, thank you.

This thesis would not exist at all without the help of Annie (for the first half) and Claire (for the second). I am enormously grateful for their supervision and advice. They are the most incredible scientists, and the most generous colleagues.

I would also like to thank Ian, for his insight, wit, and understanding. My studentship has been an amazing opportunity and has been immensely intellectually enriching. I'm very grateful.

Finally, I thank my family: Brenda, Mum, Dad, Max, and Georgia. Their love lifted burdens I thought too heavy; their support overcame hurdles I thought too high. This thesis is dedicated to them.

Preface

This thesis is a collection of the research carried out by the author in the Molecular and Population Genetics Laboratory, Wellcome Trust Centre for Human Genetics, Department of Clinical Medicine, University of Oxford, between October 2012 and April 2017, under the supervision of Professor Ian Tomlinson.

The research described herein is original, and no part has previously been submitted for a degree at this or at any other University. Unless clearly stated, all experimental and analytical work presented was conducted by the author. Notable exceptions include figure 3.6, the luciferase assays of figure 3.14, and the SMAD1/5 ChIP-seq in chapter 6, which I designed but which was conducted by Chaido Stathopoulou and Dr. James Platt. The work of others is cited with appropriate references within the text.

Aspects of the work present in chapter 3 were published in Lewis et al. (2014); the eQTL script of chapters 3 and 4 was reported in Rosmarin et al. (2014) and Cheng et al. (2015); the *GREM1* survival curves in chapter 4 were published in Davis et al. (2015).

Contents

List of Figures

List of Abbreviations

1000G 1000 Genomes project

***** stop codon

A adenine

AD autosomal dominant

AER apical ectodermal ridge

ANOVA analysis of variance

AR autosomal recessive

AUROC area under receiver operating characteristic curve

BAM binary sample alignment/map

BED browser extensible data

bp base pair

C cytosine

ChIP chromatin immunoprecipitation

chr chromosome

CI confidence interval

CNV copy number variant

CRC colorectal cancer

DNA deoxyribonucleic acid

DPD dihydropyrimidine dehydrogenase deficiency

dsDNA double-stranded DNA

E glutamate

EMAST elevated microsatellite instability at selected tetranucleotide repeats

ENCODE encyclopaedia of DNA elements

eQTL expression quantitative trait locus

ESP6500 NHLBI GO Exome Sequencing Project

G guanine

GATK genome analysis toolkit

GWAS genome-wide association study

HMPS hereditary mixed polyposis syndrome

HSD honest significant difference

IBD identity by descent

indel insertion/deletion

InSiGHT International Society for Gastrointestinal Hereditary Tumours

ISH *in situ* hybridisation

K lysine

kb kilobase pair

L leucine

LD linkage disequilibrium

LOFTEE loss of function transcript effect estimator

MAF minor allele frequency

MAP *MUTYH*-associated polyposis

MMR mismatch repair

MSI microsatellite unstable

MSS microsatellite stable

NUTS Nomenclature of Territorial Units for Statistics

OR odds ratio

P proline

PCR polymerase chain reaction

PoBI People of the British Isles

Q glutamine

R arginine

rpm rotations per minute

S serine

SNP single nucleotide polymorphism

SNV single nucleotide variant

T thymine

TCGA the cancer genome atlas

TGF-β transforming growth factor β

TMA trimethylamine

UK United Kingdom

USA United States of America

VCF variant call format

VEP variant effect predictor

W tryptophan

WHO World Health Organisation

Y tyrosine

"But, as 't is said a body, rightly mixed,
Each element in equipoise, would last
Too long and live for ever,— accordingly
Holds a germ—sand-grain weight too much i' the scale—
Ordained to get predominance one day
And so bring all to ruin and release,—
Not otherwise a fatal germ lurked here"

Robert Browning, *The Ring and the Book* (II, 137–143)

Chapter 1

Introduction

In three ways, cancer is a genetic disease. First, on any background and regardless of cause, cancer is clearly characterised by genetic mishap and derangement. Cancer cells have damaged and bizarre genomes that both allow and are caused by their breakout from normal tissue structure and function. Cancer and mutation go hand in hand.

The second sense in which cancer is genetic is more specific. Some cancers can recur within a family almost as predictably as albinism or sickle cell disease: these cancers can be described as binary Mendelian traits, inherited dominantly or recessively in predictable textbook fashion. I will refer to these heritable conditions as "syndromic" or "Mendelian" cancers. However, they are individually quite rare.

The third sense in which cancer is genetic captures those cases which do not neatly align into Mendelian Punnett squares. Unlike the rare syndromes, most cancers appear sporadically, and can therefore be modeled probabilistically. An individual's propensity to develop sporadic cancer is nonetheless heritable, albeit to a degree dependant on tumour type. Common cancer risk carries a genetic component.

This thesis is primarily concerned with the third sort of cancer genetics, the way in which, independent of lifestyle, genetic variation can influence an individual's likelihood of getting cancer. However, I will repeatedly make reference to the links between these ideas, and to the ways in which they can reinforce or illuminate each other.

The specific disease I will investigate is colorectal cancer (CRC).

1.1 Colorectal cancer

Neoplasm of the colon and rectum (ICD-10 C18–20) is the third most common cancer in the world, and the second most common in females (Stewart and Wild, 2014). In 2012, some 1.4 million CRC cases occurred (Stewart and Wild, 2014), 35,000 of which were in England (Cancer Registration Statistics, England, 2013, www.ons.gov.uk). Both incidence (shown in figure 1.1) and mortality are higher in males than in females. The disease is more prevalent in developed countries, as figure 1.2 illustrates. Indeed, there is an almost fivefold difference in CRC incidence across Europe, with the highest incidence rates in Slovakia (in men) and Denmark (in women) (Ferlay et al., 2013).

Mortality from CRC largely tracks its incidence. It is the fourth most common cause of death by cancer worldwide (Stewart and Wild, 2014). More detailed statistics exist in America, where in 2007 CRC was the third leading cause of cancer death in both sexes of all ages, accounting for just over 9% of cancer deaths (Jemal et al., 2011). Over the longer timespan of 1975–2013, this proportion was a slightly lower 8.6% of American cancer deaths (Horner et al., 2013). The NIH estimated that 49,190 Americans died of CRC in 2016 (seer.cancer.gov/statfacts/html/colorect.html).

This represents, however, a continuous and important decline in incidence and mortality for the disease in the USA. Mortality in particular has more than halved, presenting clinicians with mysteriously good news (Welch and Robertson, 2016). This decline is presumed to be the combined result of improvements in screening, surgical technique, and chemotherapy regimens. CRC is the only cancer predicted to decrease in incidence and absolute case number in America by 2030 (Rahib et al., 2014). Moreover, among some subsets of CRC patients, a total cure does seem possible. In a large international study, Renfro et al. (2015) found that "patients who remain alive and recurrence-free show a high likelihood of achieving an expected long-term survival similar to that of the general population as they reach additional year landmarks, with statistical equivalence of subsequent 3-year survival achieved at 5 years post-treatment without recurrence".

This trend is welcome, but there is more work to be done. Rahib et al. (2014) expect CRC to kill $47,000$ Americans in 2030, and America is a particular success. Survival

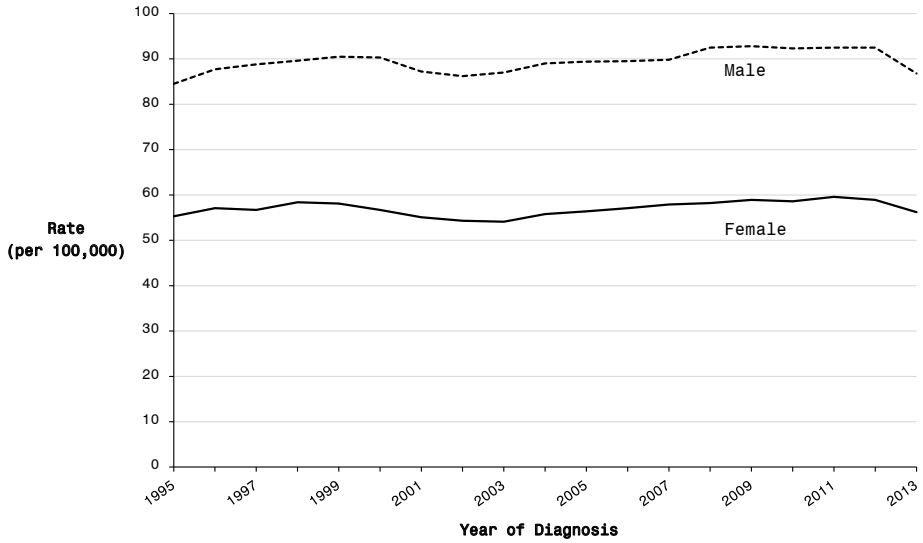


Figure 1.1: English age-standardised CRC incidence rates per 100,000 people for males (dashed line) and females (solid line). Data from www.ons.gov.uk.

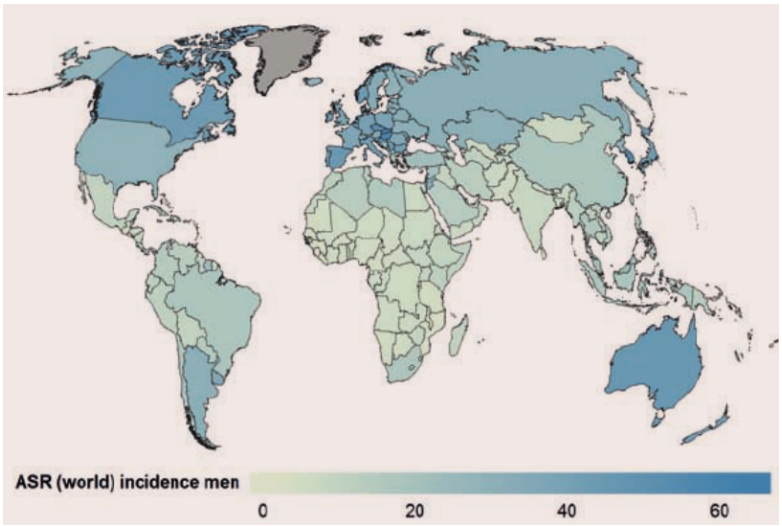


Figure 1.2: Global age-standardised CRC incidence rates per 100,000 people for males in 2012, showing bias towards countries with higher levels of development. Figure taken from Stewart and Wild (2014).

rates for CRC are better in the USA than in Europe (Gatta et al., 2000), mostly because of better screening (Gatta et al., 2003). Incidence is increasing across the developing world (Stewart and Wild, 2014). CRC is a global problem.

Some environmental CRC risk factors are known. The strongest evidence is for obesity, where meta-analyses by Okabayashi et al. (2012) and Jarvis et al. (2016) show high BMI significantly increasing the risk of colorectal adenomas and cancer, respectively. Jarvis et al. (2016) reported a CRC odds ratio of 1.23 per unit increase in BMI. The consumption of more processed or red meat and less fruits and fibre also seems to increase the relative risk of CRC (Bernstein et al., 2015). Other risk factors are so far only associated with adenomas: smoking, which roughly doubles adenoma risk (Botteri et al., 2008), and excess alcohol intake, which has a dose-dependent effect on adenoma relative risk (Zhu et al., 2014). In addition, it appears that regular exposure to aspirin can reduce CRC risk. In a large combination of five trials, daily treatment with 75mg aspirin for 5 years or longer reduced the 20-year incidence of CRC by 30% (Rothwell et al., 2010).

The assumption that simliar risks might govern adenoma and CRC incidence comes from Vogelstein et al. (1988), who first proposed the "adenoma-to-carcinoma sequence". Sequential driver mutations are thought to propel normal colon cell clones along a path from polyp to cancer. Where exactly on this path individual risk factors impinge will be discussed more in §4.3.

It is presumed that some interaction exists between environmental and genetic risk factors, though two things can be meant by "interaction". One, this could express the hypothesis that genetic risk may be felt through downstream effects, not only on DNA repair, but also on bowel transit time, alcohol consumption, BMI, microbiota, or other factors. This is plausible but difficult to test without richer outcome measures in association studies. Two, there could be formal interaction — more than additive increases in CRC risk — between different factors. Some well-powered studies have struggled to perceive these interactions (Kantor et al., 2014; Shim et al., 2016), although variants at the *HIATL1* locus do interact with alcohol consumption (Gong et al., 2016), and dietary risk also has an interactive genetic component (see §1.4.3).

Stage	Description	Substage	5-year survival
I	invasion beyond the muscularis mucosa		92%
II	invasion beyond the muscularis propria	IIA	87%
		IIB	63%
III	spread to local lymph nodes	IIIA	89%
		IIIB	69%
		IIIC	53%
IV	metastasis to distant organs and lymph nodes		11%

Table 1.1: CRC survival by AJCC substage in the SEER database (2004–2010), as given by www.cancer.org.

1.1.1 CRC subtypes

Colorectal tumours are subclassified along a bewildering array of axes. Simplest are the gross anatomical features of location and stage. But these classifications have long been supplemented by molecular information, a tendency which has accelerated in the genomics era, to provide an increasingly dominant and, for now, increasingly confusing set of subtypes.

The modern staging system is a refinement of the suggestion by Dukes (1932), who proposed four sequential categories of CRC. These are described in table 1.1, along with survival information from the American Cancer Society. NHS NICE guidelines prefer the later TNM stage classification, which is more specific about degree of invasion and node involvement (Sobin and Fleming, 1997). But in practice many clinical databases do not contain this degree of detail, and Dukes staging is used in this thesis when necessary.

Any CRC subtype can inhabit any stage, of course. It is the molecular route through the stages, and the speed of this progression, that will vary. Historically, biological differences between CRCs were attributed to left- or right-sided location in the gut; by the 1990s the preferred discriminator was the stability of DNA microsatellites.

Microsatellites are also known as simple sequence repeats. They are small (2–5bp) motifs that repeat in tandem, which are often polymorphic. The human genome is full of them; depending on definition there are at least one million microsatellite loci in the human genome, accounting for 3% of the genome (Lander et al., 2001; Warren et al., 2008). This abundance was useful for early disease mapping (see §1.2.1). In 1993, three groups independently reported a minority of sporadic CRCs which showed pervasive

change in microsatellite size between tumour and normal cells (Aaltonen et al., 1993; Ionov et al., 1993; Thibodeau et al., 1993). Their routes to this discovery were different — Aaltonen et al. (1993) were mapping Lynch syndrome, for example — and their estimates of the frequency of this microsatellite instability (MSI) ranged from 12–28% of sporadic CRC.

The causative defect in MSI CRC tumours is in the mismatch repair pathway (MMR). MMR is a form of DNA damage repair that corrects replicative errors, to which the DNA synthetic machinery is particularly prone at microsatellites. Slippage within motif repetition can either create one mismatched base, or asymmetric strand length and small insertion-deletion loops. Mismatches can become fixed point mutations if not caught before the next round of replication; loops can frameshift and truncate coding DNA. The MMR system, first discovered in *E. Coli* (Glickman and Radman, 1980; Tiraby and Fox, 1973), corrects both.

The MSI subtype, as opposed to microsatellite stable (MSS) cancers, was consistently judged to have a better prognosis, to occur in younger patients, to be more proximal in the bowel, and not to follow Vogelstein et al. (1988)'s canonical adenoma-to-carcinoma mutation sequence (Phipps et al., 2014). In particular, MSI tumours do not typically have a mutant *KRAS* gene (Ionov et al., 1993), which is frequent in MSS CRC. Nor do they display gross chromosomal instability (CIN).

CIN is a property of cancer first observed in the nineteenth century by von Hansemann (1890). The pathologist observed asymmetric or multipolar mitosis in cancer samples, and theorised that abnormal chromosome handling was either a cause or characteristic of some dedifferentiating drive within tumour *Mutterzellen* (stem cell analogues). Theodor Boveri — who is often unfairly credited with von Hansemann's ideas (Bignold et al., 2006) — corroborated this theory with detailed descriptions of the chromosomally unusual sea urchin cells formed by the simultaneous fertilisation of an egg by two sperm (Boveri, 1907). These double fertilisations produced supernumerary centrosomes, an immediate split into not two but three or four daughter cells, a random inheritance of chromosome number and type, and subsequent death at gastrulation. Boveri argued that this meant that each chromosome was responsible for different as-

pects of heredity, and moreover observed that, instead of premature death, on occasion a larval cell endowed with an unsual chromosomal complement would display *schrankelose Vermehrung* ("unlimited growth", Boveri, 1914).

The majority of CRCs are aneuploid: they have an abnormal chromosome number, due to CIN (Thompson and Compton, 2008). This is mostly due not to mitotic dysfunction, but rather to defective replication forming misshapen chromosomes which then missegregate (Burrell et al., 2013). Recurrent changes in CRC include gain of chromosomes 1q, 7, 8, 12q, 19q, and 20; and loss of chromosomes 18 (where *SMAD4* resides) and 17p (containing *TP53*; Muzny et al., 2012).

The mututal exclusivity of MSI and CIN provided the simplest and most clear-cut molecular categories of CRC until the genomics era. An additional trait, called CpG island methylator phenotype (CIMP), was often also included in classifiers, but the usefulness of this information from a purely biological standpoint was moot. CIMP is an overabundance of gene silencing by promoter methylation; almost all sporadic MSI tumours have the MMR system curtailed by CIMP causing methylation at *MLH1* (Goel et al., 2007). But this is a one-way relationship. Most MSI tumours have CIMP, but not all CIMP tumours are MSI (Samowitz et al., 2005). Therefore, the phenotype affords modest additional discriminatory power.

The 2010s have seen a renewed enthusiasm for subclassification. These efforts are mostly based on the clustering of expression data from microarrays or RNA sequencing to produce molecular CRC subtypes. In some cohorts this produced three categories (Roepman et al., 2013), in others four (Perez-Villamil et al., 2012), five (Budinska et al., 2013; Schlicker et al., 2012), or even six (Marisa et al., 2013; Sadanandam et al., 2013). The biological and prognostic relevance of these unsupervised clusters is often difficult to fathom. To unify the proliferating and competing systems, Guinney et al. (2015) formed an international consortium to define the consensus molecular subtypes (CMS) of CRC.

The CMS system is one of the two current molecular classifications of CRC. The other is based on DNA sequencing from the cancer genome atlas (TGCA, Muzny et al., 2012). This is based on somatic mutation pattern, and distinguishes three groups:

microsatellite stable CIN tumours; MSI tumours; and a small group of "ultramutant" cancers with dysfunctional *POLE*. The TCGA and CMS categories are summarised in table 1.2.

TCGA group	Proportion	CMS group	Description	Proportion
MSI	13%	CMS1	MSI/immune	14%
CIN/MSS	84%	CMS2	canonical/epithelial	37%
		CMS3	metabolic	13%
		CMS4	mesenchymal	23%
Ultramutant	3%			

Table 1.2: Current molecular subtypes of CRC, from Muzny et al. (2012) and Guinney et al. (2015). The strongest agreement is on the proportion and prognosis of the MSI group.

Ironically, after a decade of deeper digging, the best classifier remains the presence of MSI, although recent work also confirms a distinct clinical behaviour of *POLE*-mutant CRC (Domingo et al., 2016). The biology of the three MSS CMS groups (CMS2–4) needs more exploration.

1.2 Mendelian CRC

Known Mendelian syndromes make up approximately 5% of CRC incidence (Jasperson et al., 2010; Rustgi, 2007). This estimate is mostly a count of Lynch syndrome patients, who suffer this most common CRC syndrome (see §6.1.1). Second most common is familial adenomatous polyposis (FAP). As well as these, there is a long tail of very rare conditions including *MUTYH*-associated polyposis (MAP), *NTHL1*-associated polyposis (NAP), polymerase proofreading-associated polyposis (PPAP), serrated polyposis syndrome (SPS, previously known as hyperplastic polyposis), hereditary mixed polyposis syndrome (HMPS), juvenile polyposis syndrome (JPS), and Peutz-Jeghers syndrome.

It is helpful to follow Tomlinson (2015) in grouping these syndromes by the colorectal lesion to which they predispose patients. Most appear to have polyposis as a primary endpoint: only Lynch syndrome is associated with increased CRC risk in the absence of polyps. The Mendelian conditions can then be subdivided by polyp type. FAP, MAP, NAP and PPAP predispose to classical adenomatous polyps; JPS and Peutz-Jeghers predispose to hamartomatous polyps; SPS causes serrated-type polyps; and HMPS leads

Syndrome	Gene(s)	Inheritance	Proportion of CRC
Lynch	*MSH2*, *MSH6*, *MLH1*, *PMS2*	AD	3–4%
FAP	*APC*	AD	1%
MAP	*MUTYH*	AR	$< 1\%$
NAP	*NTHL1*	AR	$< 1\%$
PPAP	*POLE*	AD	$< 1\%$
Peutz-Jeghers	*STK11*	AD	$< 1\%$
JPS	*BMPR1A*, *SMAD4*	AD	$< 1\%$
HMPS	*GREM1*	AD	$< 1\%$
SPS	-	AD	-

Table 1.3: Summary of Mendelian CRC or polyposis syndromes. AD, autosomal dominant; AR, autosomal recessive.

to a mixed polyposis of all types. See table 1.3.

The risk of disease conferred by mutation in Mendelian syndrome genes is variable, modified by additional variants or by environmental factors. Reeves et al. (2008) found that an *IGFR1* promoter CA dinucleotide repeat predicted age of disease onset in both Polish and Austrian Lynch syndrome patients, independent of inherited mismatch mutation, gender, and family. Variants in *MTHFR*, an enzyme involved in folate metabolism, are protective in Lynch syndrome (Pande et al., 2007; Reeves et al., 2009). And homozygosity at an *HFE* variant, causing hereditary haemochromatosis, significantly correlated with absolute risk and earlier age of onset of CRC in Lynch patients (Shi et al., 2009). There is no evidence, however, that variants associated with sporadic CRC risk influence Lynch syndrome penetrance or outcome (Win et al., 2013).

The causative gene or genes have been discovered for all but one of the conditions above (SPS; see table 1.3). I will not review every syndrome, but rather will provide an overview of the genetics of JPS, HMPS, and PPAP. This subset should adequately illustrate both the past decade's advances in disease mapping, and some shared biological characteristics of the diseases. The earliest attempts to locate pathogenic germline genetics took advantage of linkage.

1.2.1 Linkage

The intellectual apparatus of linkage studies emerged alongside the discipline of genetics itself, during the fervid scientific activity of the *Belle Époque*. In 1900, after Darwin, but before Bateson coined the phrase "genetics" in 1905, and long before a truly molecular account of heredity, a handful of independent inquirers either rediscovered or recreated the work of Gregor Mendel (Correns, 1900; Mendel, 1865). The codification of inheritance, and specifically Mendel's law of independent assortment, spurred great debate as to the location and nature of genes. Did they reside on chromosomes, as suggested by Sutton (1903) and Boveri (1904)? Thomas Hunt Morgan thought this idea problematic:

> *"Since the number of chromosomes is relatively small and the characters of the individual are very numerous, it follows on the theory that many characters must be contained in the same chromosome. Consequently, many characters must Mendelize together. Do the facts conform to this requisite of the hypothesis? It seems to me that they do not."*
>
> Morgan (1910a)

And yet Morgan knew that Mendelian phenotypic distributions were inadequate biological descriptors, too. Non-Mendelian assortment existed and had been reported. Bateson et al. (1905) observed a mysterious coupling of petal colour and seed shape in the sweet pea; in the same year as Morgan's critique of chromosome theory above, he and his colleagues described a trait — white eyes in fruit flies — whose inheritance pattern was sex-linked. The heredity of white-eye was parsimoniously explained by its coupling to a mutant X chromosome, a hypothesis confirmed by breeding experiments (Morgan, 1910b). But this explanation strongly suggested that the X carried information.

Two additional lines of investigation strengthened the case for the chromosomal theory. One, the recently emerging understanding of the mechanics of germ cells and meiosis, was furnished by Oscar Hertwig and August Weismann in the late nineteenth century. The other was the meticulous descriptions by Janssens of what he termed "chiasmatypie": the twisting, merging, and "otherwise completely inexplicable interweavings" of chromatids during meiosis (Janssens, 1909). This was the first account of chromosomal crossover. Morgan realised that chiasmatypie might void his previous objections to chromosomal heredity, and in 1911 provided the beautful and pithy theory

of genetic linkage.

"[Observations of coupling] *are a simple mechanical result of the location of the materials on chromosomes, and of the method of union of homologous chromosomes* [...] *Instead of random segregation in Mendel's sense we find "association of factors" that are located near together in the chromosomes. Cytology furnishes the mechanism that the experimental evidence demands."*

Morgan (1911)

Characteristics on one chromosome are prevented from always "Mendelizing together" by crossover during meiosis. The physical reality of this genetic exchange during recombination was not confirmed until the 1930s, by Creighton and McClintock (1931) in maize and by Stern (1931) in fly a few weeks later (to say nothing of the mystery of the molecular basis of this genetic information). And it would take an additional fifty years before Morgan's theory could serve medical genetics, via its obvious corrolary: that those regions of the genome which cosegregate with heritable disease must lie close to pathogenic variation. By following variable genetic elements in pedigrees, causal regions can be discovered.

By convention, evidence for linkage is quantified by the log of odds (LOD) score, which is log_{10} of the odds of observed linkage not occurring by chance. Odds of at least 1000 to 1 ($LOD = 3$) are considered adequate to infer linkage between two traits.

The three diseases discussed below track the changing variants used in linkage mapping as technology improved. In the 80s, restriction digestion and electrophoresis were used to describe restriction fragment length polymorphisms (Botstein et al., 1980). This led to the first disease mapping, of Huntington's to chromosome 4 (Gusella et al., 1983). Later, the invention of PCR allowed a simpler amplification of polymorphic regions, with primers flanking microsatellites.

1.2.2 JPS

After an exploratory study ruled out *PTEN* and *MCC* germline mutation in JPS (Leggett et al., 1993), Howe et al. (1998a) investigated twelve other loci and successfully mapped JPS. The authors used simple sequence repeat polymorphisms as markers, looking for cosegregation in a large JPS kindred. Only at 18q21.1 was marker $LOD > 3.0$,

surrounding the genes *DCC* and *SMAD4*. Later that year Howe et al. (1998b) pinpointed the kindred's causative mutation, a frameshift deletion in *SMAD4*.

However, Howe et al. (1998b) found no mutations in sequenced *SMAD4* exons in four unrelated patients. Also, Houlston et al. (1998) saw no linkage between *SMAD4* and JPS in eight affected families and found only one instance of *SMAD4* mutation in a larger patient set using conformation-specific gel electrophoresis. Though *SMAD4* mutations caused JPS in some cases, they did not underlie every instance of the disease.

Howe et al. (2001) performed a genome-wide linkage screen with simple sequence repeat polymorphisms in four families with JPS but without *SMAD4* mutation and observed high *LOD* scores around chromosome 10q22–23. As *PTEN* mutations had been excluded, this left *BMPR1A* as the most interesting gene at the *LOD* peak. Custom microsatellite markers flanking the gene showed total disease cosegregation, with a recombination factor of 0. Howe et al. (2001) proceeded to Sanger sequence every exon of *BMPR1A*, and found four separate deleterious mutations in each of the four families, every one cosegregating with disease.

BMPR1A is a type I bone morphogenetic protein (BMP) receptor kinase. In partnership with an activating type II BMP receptor, BMPR1A phosphorylates the receptor-regulated SMAD proteins SMAD1, SMAD5, and SMAD9. These then oligomerise with the co-SMAD, SMAD4, and move to the nucleus to act as transcription factors. Figure 1.3 diagrams this signalling pathway.

Therefore, JPS is caused by mutation of different components of one signal transduction cascade: the BMP pathway. Mutations in both the BMP receptor and its activated transcription cofactor are sufficient to cause polyposis.

1.2.3 HMPS

HMPS was formally described by Whitelaw et al. (1997), who did not exclude the possibility that it was an unusual form of JPS. A large Ashkenazi family (SM96) suffered dominantly inherited early-onset CRC with mixed polyposis. Linkage of the disease to *APC*, *TP53*, *MSH2*, *MLH1*, and *DCC* (and hence *SMAD4*) was ruled out by targeted mapping of restriction fragment length and simple sequence repeat polymorphisms. A

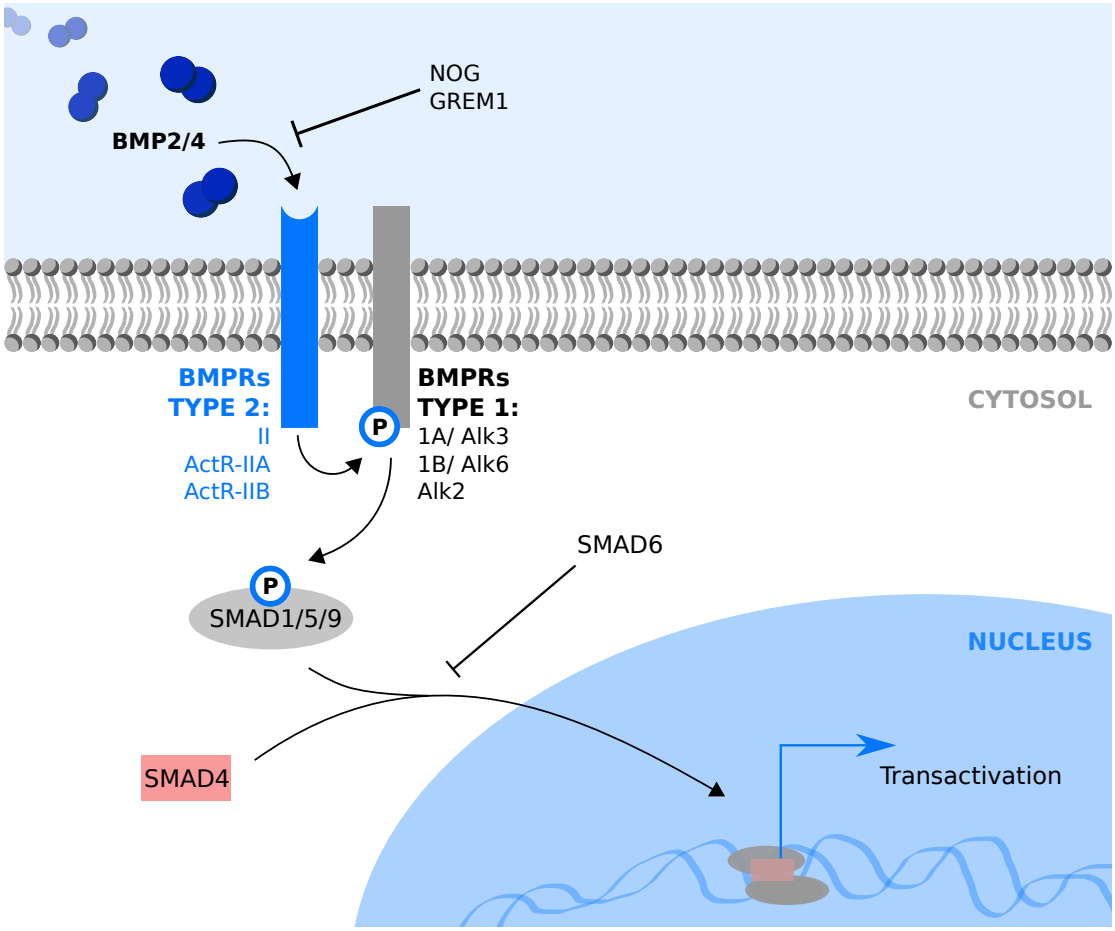


Figure 1.3: Simplified diagram of the BMP pathway. Circled Ps represent phosphorylation steps. SMAD4 and SMAD1/5/9 together mediate the BMP suppression of intestinal Lgr5^{+} stem cell renewal, among other effects (Qi et al., 2017).

year prior, a genome-wide linkage study in the same kindred had claimed chromosome 6q as the pathogenic locus (Thomas et al., 1996). This later proved incorrect, however, when an individual from SM96 without the disease-associated 6q haplotype developed multiple colorectal adenomas at an early age (Jaeger et al., 2003).

The first correct HMPS mapping is therefore a matter of semantics. Tomlinson et al. (1999) correctly identified chromosome 15q14–q22 in a genome-wide linkage study, hypothesising a local *CRAC1* (colorectal adenoma and carcinoma 1) gene. But this was in a separate family not yet diagnosed with HMPS. Jaeger et al. (2003) rejected the chromosome 6 mapping, repeated a genome-wide microsatellite linkage study, and found strong evidence for a 15q13–q14 HMPS locus in SM96, the CRAC1 family, and a third family of Ashkenazim with multiple adenomas. All affected patients in each kindred shared the same haplotype.

Due to its highly unusual nature, the underlying pathogenic change at the HMPS locus was not understood until 2012. After finding no novel or pathogenic changes in the three genes at the locus (*GREM1*, *SCG5* and *FMN1*) by sequencing, Jaeger et al. (2012) designed a custom oligonucleotide array to assess local copy number. The mutation was found. All HMPS patients were heterozygous carriers of a 40kb duplication, incorporating the last four exons of *SCG5* and ending just upstream of *GREM1*. Somehow, this duplication vastly increases the expression of *GREM1*, ectopically. It shows perfect disease cosegregation.

GREM1 is an extracellular BMP antagonist (see figure 1.3) discussed in greater detail in chapter 3. It is the third member of the canonical BMP signalling pathway to be directly implicated in Mendelian CRC.

The Jaeger et al. (2012) duplication is only found in Ashkenazim and is almost certainly a founder mutation that has spread throughout a single, dispersed kindred. Two other reports describe Ashkenazi patients with this same *GREM1* duplication (Plesec et al., 2016; Ziai et al., 2016). Additionally, Rohlin et al. (2015) found a second, 16kb duplication upstream of *GREM1*, in a family beset by CRC and a polyposis of possibly mixed type. This duplication was found by multiplex ligation-dependent probe amplification following a linkage study on thinned exome sequencing, which highlighted

the *GREM1* 5′ untranslated region. Rohlin et al. (2015) reported no Ashkenazi heritage for this kindred, where the smaller duplication was found exclusively in the four sufferers out of eight family members interrogated. As in the Ashkenazi duplication, gut *GREM1* expression was increased, although to a lesser extent.

1.2.4 PPAP

Research into JPS and HMPS genetics spans technology that began with the very first, low-coverage studies and led to exome sequencing; it's only appropriate that PPAP should bring us up to date by using whole-genome sequencing (WGS) to discover Mendelian mutations — without entirely abandoning older disease mapping techniques.

WGS variant discovery often relies on aggressive filtering strategies. An overall list of variation is subset on multiple parameters to create a tractable candidate list. For example, Palles et al. (2013) searched for coding or splice site variants shared among 4 of 13 families with multiple, early adenomas. They found none, however, and so turned to preexisting linkage data within two of the families.

It was this intersection between shared rare variants from WGS and shared regions from linkage that allowed the discovery of each of these families' pathological mutations. One had a missense variant in *POLE*, the other a missense variant in *POLD1*. These genes encode the catalytic and proofreading subunits of DNA replication polymerases ϵ and δ, respectively. The mutations disrupt the exonuclease (proofreading and mismatch-repairing) domain of the proteins and cause dominantly inherited adenomas and early or multiple CRCs. *POLD1* mutations also predispose to endometrial cancer in females. Briggs and Tomlinson (2013) named the condition PPAP, and other groups have subsequently reported cases (Elsayed et al., 2014; Goodman, 2014; Valle et al., 2014), though the condition accounts for less than one percent of early-onset CRC cases in all studies and is therefore rare in the general population.

The cancer syndromes discussed illustrate four principles. The first is typological: much of the success of these endeavors requires tight and valid biological categorisation. Otherwise one ends up with some of the confusion seen in the early mapping of HMP-S/CRAC1. FAP, which I did not discuss, is an end-stage agglomeration of previously

distinct traits, subsuming "APC" and "Gardner's syndrome". Nor is the boundary between HMPS and JPS phenotypically clear-cut, as the exchange between Cheah et al. (2012) and Tomlinson et al. (2012) illustrates. But the genetics are unambiguous and should be considered pathognomonic. A purely genetic classification of Mendelian disease is vastly preferable and — given the apparent takeover of genetics-led rather than descriptive disease discovery, as in PPAP — may now be inevitable in any case.

Second, CRC resembles other cancers in how its Mendelian forms corroborate Knudson's most famous idea. Knudson (1971) proposed that germline and somatic mutations in the same gene might both cause retinoblastoma, though he did not explicitly generalise to all cancer syndromes. Nevertheless this remarkable relationship between the rare and the common disease has repeatedly resurfaced. The causal genes in FAP (Solomon et al., 1987), Lynch syndrome (Mensenkamp et al., 2014), and PPAP are all mutated in sporadic CRC. Indeed, *POLE* mutation, though uncommon, defines a mismatch-repair proficient, immunogenic, hypermutant CRC tumour subset with an excellent prognosis (Domingo et al., 2016).

The third principle is technical: disease mapping charts the course between PCR and WGS. The extremely rapid discovery of PPAP illustrates the power of modern methods. It's worth observing that the hiatus in Mendelian enquiry at the turn of the millenium in the narratives above reflects a broader preoccupation of the genetics community with other large projects at the time, and a turning away from rare diseases more generally (see §1.4). Available technology has both limited and directed the fashions in genetic research.

This leads me to the fourth principle, somewhat at odds with Knudson's insight. Most CRC is not Mendelian. It's not random, either: some degree of CRC risk is heritable. And it's not clear the extent to which Mendelian genetics can illuminate this risk. It's worth being pedantic here. A loose statement of the two-hit hypothesis is that inherited and somatic mutations in the same gene will produce the same cancer. This neither supports nor precludes the idea that germline variation affecting an entirely separate set of genes might govern CRC *risk*. This possibility, alongside the individual rarity of the syndromes and emerging array technology, spurred inquiry into the subtler

genetics of CRC predisposition through the genome-wide association study (GWAS). But before one can discover the genetics of heritable cancer risk, one must gain a sense of how much there is to explain — a measurement of its heritability.

1.3 The heritability of complex traits

"For, surely, no one fails to see that the appearance and habits, and generally, the carriage and gestures of children are derived from their parents. This would not be the case if the characteristics of children were determined, not by the natural power of heredity, but by the phases of the moon and by the condition of the sky."

Cicero, *De Divinatione*, 2.94

Cicero's attribution in 44 B.C. of Chaldean astrological nonsense to certain Stoic thinkers is one of the first records of philosophical concern with twins' mutual resemblance. But what he means by *vis et natura gignentium* is unclear, invoking and muddying to modern ears that tired dichotomy of nature versus nurture, genes versus environment. We have had at our disposal for decades now an easy bypass of the issue, which is to accept both. "The concept of *heritability* replaced a meaningless discussion of whether heredity or environment is all-important" (Crow, 1971, emphasis mine).

Heritability is different for different traits, and can be measured. Mendelian traits are monogenic and highly penetrant; continuous and complex traits are not, but can be modeled as the sum of many small effects. As Fisher stated in 1918 (while offhandedly inventing the statistical definition of variance), "the simplest hypothesis, and the one which we shall examine, is that such features as stature are determined by a large number of Mendelian factors". This hypothesis, he concluded, "seems to fit the facts very accurately" (Fisher, 1918).

Heritability is the proportion of phenotype variance (σ^2_P) attributable to genetic variance (σ^2_G). The definition of the second term is key. If we are concerned with the similarity bewteen parent and offspring, who can share a maximum of one allele per locus, then genetic variance refers to the additive effect of these alleles. The ratio of additive variance (σ^2_A) over phenotypic variance is the "narrow-sense" heritability, h^2

(equation 1.1).

$$h_2 = \frac{\sigma_A^2}{\sigma_P^2} \tag{1.1}$$

h^2 is commonly derived from twin studies using the "ACE" model, splitting phenotypic variance into three components: A for additive genetic input, C for environmental inputs common to twins, and E for private environmental inputs. If r_{MZ} is monozygotic phenotypic correlation, and r_{DZ} is dizygotic correlation, then

$$h^2_{ACE} = 2(r_{MZ}r_{DZ}) \tag{1.2}$$

Narrow-sense heritability is only one component of total genetic variance, as it does not model interactions between loci (epistasis) and alleles (dominance). These are genotypic rather than allelic factors and variance in these parameters, as well as maternal and paternal effects, all contribute to total genetic variation (equation 1.3, where σ_D^2 and σ_I^2 are the partitions of genetic variance due to dominant effects and epistasis, respectively).

$$\sigma_G^2 = \sigma_A^2 + \sigma_D^2 + \sigma_I^2 \tag{1.3}$$

The ratio of σ_G^2 over σ_P^2 is the "broad-sense" heritability, H^2 (Dempster and Lerner, 1950). This partitions out of σ_P^2 any environmental or covariant contributions.

When we are concerned with a binary phenotype, such as CRC case or control status, rather than a spread of values, such as height, to what does σ_P^2 refer? There are two ways of modelling dichotomous phenotypic variance. The first is to simply treat the trait as a binary variable, in which case its variance will be $K(K-1)$, where K is the probability or incidence of the trait. This is undesirable, as it makes comparisons across traits with different incidences difficult, and can obscure truly additive contributions of genetic variance (Dempster and Lerner, 1950). Following earlier work by Wright (1926), Lush et al. (1948) were the first to apply the second model of dichotomous phenotype variance. This "threshold" model treats the trait as a normally distributed liability, the sum of normally distributed environmental and genetic components. Disease manifests when this liability passes a certain threshold value.

1.3.1 Heritability of CRC

That sporadic CRC incidence has a genetic component is well founded. An individual with at least one first-degree relative with CRC has a relative risk of the disease of 2.25; their relative risk remains roughly double that of the normal population if that first-degree relative merely has adenomas (Johns and Houlston, 2001). As is surprisingly common across traits, however, exact estimates for the heritability of CRC are few and uncertain.

Early twin studies were small but reasonably consistent. Comparing monozygotic to dizygotic twin concordance rates allowed Verkasalo et al. (1999) to estimate a CRC additive genetic component explaining 35% of phenotypic variation; Ahlbom et al. (1997) gave a higher figure of 43% in males only (they had no informative female twins). Lichtenstein et al. (2000) used a combination of Scandinavian twin registries, including both the Verkasalo et al. (1999) and the Ahlbom et al. (1997) cohorts, to amass some 1300 twins where at least one sibling had CRC. Assuming a total correlation between additive genetic and shared environmental components of phenotypic variance in monozygotic twins, they generated an estimate of the narrow-sense CRC heritability of 35%. This estimated heritable component was slighly higher for CRC occurring before age 75.

Other estimates of CRC heritability are derived from modelling concordance between many different types of relatives in the large Swedish Family Cancer Database (Czene et al., 2002; Hemminki et al., 2001). The database contains millions of individuals. For each type of sex-specific relationship (father-daughter, sister-sister, *etc.*), concordance can be quantified by tetrachoric correlation of contingency tables, and the liability threshold separately estimated from incidence (as explained in Lee et al., 2011). In the most recent study, there were 860 CRC-concordant relatives, and many more disparate pairs. Czene et al. (2002) produced considerably lower estimates than the earlier twins: 12–18% for colon cancer, and 8–13% for rectal cancer narrow-sense heritability.

At first glance these estimates seem discomfortingly conflicting. But in fact their 95% confidence intervals, when reported, overlap. Lichtenstein et al. (2000) report an additive heritability of 10–48%, which encompasses the results of Czene et al. (2002). Nevertheless, additional updates to this figure would be welcome, especially given much

of the confused rhetoric around heritability, as discussed in chapter 7.

Having established the reality of a genetic component in sporadic CRC risk, and armed with an imperfect estimate of the magnitude of that contribution, the task is now to assess what particular biological changes actually underlie the disease, using GWAS.

1.4 GWAS

1.4.1 Theory and early practice

GWAS is a complex way to answer a simple question. For a heritable trait, it interrogates the genetic differences between those with and without the trait. An assumption of GWAS is that, so long as their risk has a respectable genetic component, common diseases such as sporadic CRC will be partially caused by common variation. Therefore, GWAS examines common variant allele frequency differences between large numbers of cases and controls.

GWAS is unlike linkage studies in that it requires no relationship information (indeed, relatives should not be used) and looks for correlation rather than cosegregation. The resolution of a linkage study is limited by the number of recombination events in kindreds; the resolution of GWAS is limited by the number of variants genotyped and by the linkage disequilibrium (LD) structure of the population in question.

There are two immediate obstacles to successful GWAS. The first is genotyping technology. When GWAS was in its infancy, the cost of WGS remained prohibitively expensive. Even testing most common variants was effectively impossible. This issue is exacerbated by the second obstacle: GWAS deals with very large numbers. For an effective study, very many variants must be genotyped in very many people. This further increases the cost of GWAS, and also introduces a considerable risk of false-positive findings. Moreover, proximate common variants have correlated allele frequencies due to LD, as Morgan (1911) predicted.

The early solution to these problems was ingenious. An immediate extension of the Human Genome Project (Lander et al., 2001), the International Hapmap Project investi-

gated the haplotype structure of four separate human populations: Yoruba from Ibadan, Nigeria; Japanese inhabitants of Tokyo; Han Chinese from Beijing; and Americans with Northwestern European ancestry from Utah (The International Hapmap Consortium et al., 2003). As well as a tool for studying recent human evolutionary history, the project was necessary for GWAS. Not only did it expand the cataloguing of variation within these populations, especially of single nucleotide polymorphisms (SNPs), but the degree of correlation between these variations enabled the first maps of recombination landscape of the human genome (The International HapMap Consortium, 2005). Recombination rate is not constant over each nucleotide of the genome (Jeffreys et al., 2001; Myers et al., 2005), nor is the work of recombination "done", in the sense of having had enough time to completely shuffle the deck of human genetics. As such, human populations exhibit contiguous, coinherited haplotypes that can be "tagged" by a single SNP lying within them. African haplotypes are shorter and more numerous, as Africa housed the older founder population out of which all other human populations arose (Ramachandran et al., 2005; Reich et al., 2001; Shifman et al., 2003). Early GWAS therefore only typed and tested a subset of the $\sim$ 10 million common SNPs then catalogued, a 500,000-strong subset that represented LD blocks across the nonrepetitive genome. In this way, the problem of multiple testing was reduced by two orders of magnitude, and contemporary array technology was competent to the task.

Ozaki et al. (2002) carried out the first GWAS, for myocardial infarction, having successfully typed just 65,671 SNPs in 95 cases and 658 controls. Their published association was not corrected for multiple testing, and would not be considered a true result today. It is now standard to correct association P values by the overly conservative Bonferroni method (Johnson et al., 2010). This means that a genome-wide significant association requires $P < 5 \times 10^{-8}$, assuming 1,000,000 independent tests. The corrected value for Ozaki et al. (2002) would be $P = 1$. As a result, the study is often rather unfairly forgotten in favour of the first *successful* GWAS, associating macular degeneration with variation at the *CFH* gene (Klein et al., 2005).

These first, dripwise attempts were harbingers of a great GWAS flood. As genotyping technology cheapened, reports proliferated exponentially. Topol et al. (2007)

declared the advent of "the genomics gold rush": in 2010 over half of the articles in Nature Genetics were GWASs (Ikegawa, 2012). But early GWAS was underpowered, and had too low a resolution, especially when attempting to refine an association hit. Ever-increasing case-control cohort sizes helped the first problem. Genotype imputation helped the second.

1.4.2 Imputation

The logic of GWAS array design is that, knowing something of the LD stucture of the genome, individual SNPs can stand as proxies for their surrounding, correlated, variants. The corollary is that missing genotypes must be to some degree inferable from surrounding SNPs. Additional genetic information can be imputed (Howie et al., 2009; Marchini et al., 2007).

Imputation requires reference haplotypes, which will supply genotypes to the sparser input and which therefore must come from a similar population. These can be explicitly phased, or estimated. This reference information was part of an updated HapMap (The International Hapmap Consortium, 2010), which was then superseded by its successor, the 1000 Genomes project (Abecasis et al., 2010). The 1000 Genomes project (1000G) catalogues more rare variants in more populations, and is a common imputation backbone (Wood et al., 2013). Increasingly, however, these shared resources are supplemented or replaced by private whole-genome sequencing (WGS), as in this thesis.

Modern methods estimate the phase of all typed SNPs in both reference and study populations using iterations of a Markov chain Monte Carlo (MCMC) algorithm, and then impute genotypes in the sparse study set (Howie et al., 2009). At the end of this iterative cycling, missing genotypes are replaced by allele probabilities, with information scores for each locus.

The great strength of imputing missing genotypes is that it allows the combined analysis of samples originally typed on different arrays. As such, it was a crucial technology for the expansion of GWAS, because multiple small cohorts could be pooled to provide greater statistical power. Imputation also enables the "fine-mapping" of trait-associated loci by increasing spatial resolution, potentially improving the best tagging

variant, or even highlighting the previously untyped causal variation itself. This second possibility marks a discontinuity in the history of disease mapping: with imputation it is possible (although still unlikely) to simultaneously map a disease locus and to discover its causal variant.

Cheaper arrays, large public databases of variation, and imputation have contributed to the enormous expansion of GWAS. As of March 2017, over 3600 GWAS papers have been entered into the NHGRI-EBI GWAS catalogue (www.ebi.ac.uk/gwas), representing some 36,500 associations with hundreds of traits.

1.4.3 CRC GWASs

GWAS has implicated at least 30 loci in hereditary CRC risk. Table 1.4 summarises the state of the art, although it is merely a list of loci associated with CRC in Europeans, and is therefore missing two kinds of information. First, since this thesis deals exclusively with European genetics, I have not reported loci found in Asian-only CRC GWASs that have not yet been replicated in Europeans (see, for example Jia et al., 2012; Zeng et al., 2016; Zhang et al., 2014a,b). The table shows which signals have been validated (or discovered) in multiple populations, which for the moment essentially means in both Europeans and Asians, there being only one published CRC GWAS in another population (the small African American set in Kupfer et al., 2014). Second, table 1.4 only shows loci, not individual risk signals. Some loci harbour more than one risk-associated variant, whether causal or tagging. These include 10q24.2 (*ABCC2*, or possibly *ENTPD7*), 12q13.13 (*DIP2B*), 14q22.2 (*BMP4*), 20p12.3 (*BMP2*), and, as discussed at length in chapter 3, the *GREM1* locus, 15q13.3. So far only two CRC GWAS loci have been shown to interact with environmental risk factors: the *EIF3H* locus SNP rs16892766, with vegetable consumption (Hutter et al., 2012), and the *GATA3* locus SNP rs4143094, with processed meat consumption (Figueiredo et al., 2014).

It's worth highlighting two important points. One, I am continuing an ignoble tradition in supplying genes for every risk association, as though that were the end of the matter. It's almost professional convention at this point to provide a nearby biological effector for reported GWAS loci or SNPs. But, especially in older papers, there

Locus	Gene	Reference	Multi-ethnic?
1p36.22	*MTHFR*	Hubner and Houlston (2007)	
1p36.12	*WNT4/CDC42*	Al-Tassan et al. (2015)	
1q25.3	*LAMC1*	Kinnersley et al. (2012)	
1q41	*DUSP10*	Houlston et al. (2010)	Jia et al. (2012); Zhang et al. (2014a)
3p22.1	*CTNNB1*	Schumacher et al. (2015)	
3p14.1	*LRIG1*	Schumacher et al. (2015)	
3q26.2	*MYNN/TERC*	Houlston et al. (2010)	Jia et al. (2012)
5p15.33	*TERT*	Kinnersley et al. (2012)	
6p21	*CDKN1A*	Dunlop et al. (2012)	Zhang et al. (2014a)
8q23.3	*EIF3H*	Tomlinson et al. (2008)	Zeng et al. (2016)
8q24.21	*POU5F1P1*	Tomlinson et al. (2007)	Jia et al. (2012)
10p14	*RNA5SP299*	Tomlinson et al. (2008)	Jia et al. (2012); Zhang et al. (2014a)
10p13	*ERAP1*	Al-Tassan et al. (2015)	
10q25	*VTI1A*	Wang et al. (2014a)	yes, by design
10q24.2	*ABCC2*	Whiffin et al. (2014)	
11q13.4	*POLD3*	Dunlop et al. (2012)	Jia et al. (2012)
11q23	*POU2AF1*	Tenesa et al. (2008)	Zhang et al. (2014a)
12p13.32	*CCND2*	Whiffin et al. (2014)	Zhang et al. (2014b)
12q13.13	*DIP2B*	Houlston et al. (2010)	Zhang et al. (2014b)
12q24.12	*SH2B3*	Timofeeva et al. (2015)	
12q24.22	*NOS1*	Schumacher et al. (2015)	
14q22.2	*BMP4*	Houlston et al. (2008)	Jia et al. (2012)
15q13.3	*GREM1*	Jaeger et al. (2008)	Kupfer et al. (2014)
16q22.1	*CDH1*	Houlston et al. (2008)	Kupfer et al. (2014)
16p13.2	*C16orf72*	Al-Tassan et al. (2015)	
18q21.2	*SMAD7*	Broderick et al. (2007)	Jia et al. (2012); Zhang et al. (2014a)
19q13.1	*RHPN2*	Houlston et al. (2008)	Jia et al. (2012); Zhang et al. (2014a)
20p12.3	*BMP2*	Houlston et al. (2008)	Zhang et al. (2014a)
20q13.13	*PREX1*	Schumacher et al. (2015)	Zeng et al. (2016)
20q13.33	*LAMA5*	Houlston et al. (2010)	Zhang et al. (2014a)
Xp22.2	*SHROOM2*	Dunlop et al. (2012)	

Table 1.4: Summary of all reported genome-wide significant European CRC GWAS loci as of March 2017. Only the first significant report of each locus is cited; details on fine-mapping and splitting of the signals at various loci are provided in later sections. Genes reported are often merely the nearest or most plausible candidates and should be viewed with suspicion. "Multi-ethnic" means associated in Europeans and either Asians or African Americans.

is almost no evidence linking variants to genes beyond simple proximity. Even that is sometimes not clear-cut: the CRC risk SNPs rs4813802 and rs961253 are roughly 50 and 350 *kilobases* away from *BMP2*, respectively. Genes are mostly reported to add interest to a GWAS paper. But some association peaks really do lie in gene deserts (the 10p14 signal, for example), and anyhow the assumption of a variant enhancer element allele-specifically regulating its nearest gene does not fit our understanding of genetic regulation: enhancers can and do leapfrog their neighbouring genes to affect more distant partners (Pregizer and Mortlock, 2009). Increasingly, bioinformatic databases of correlations between variants and gene expression are able to identify the molecular effect of GWAS associations with more finesse (The GTEx Consortium et al., 2015). However, many older GWAS hits still lack any evidence of a functional link to their reported target gene.

Despite that caveat, I can make a second point. There are groups of genes in the list that are involved in similar processes: cell cycle regulators (*CDC42*, *CCND2*, *CDKN1A*, *DUSP10*), telomeric proteins (*MYNN*, *TERT*, *TERC*), cell adhesion molecules (*PREX1*, *CDH1*, *SHROOM2*). These are biologically sensible groups of cancer-driving genes. Moreover, some of these groups functionally recall the Mendelian cancer syndromes. Variation around *POLD3* increases the risk of sporadic CRC just as *POLD1* mutation causes hereditary cancer; multiple members of the BMP pathway — *BMP2*, *BMP4*, *LAMA5*, and *GREM1* — have CRC risk SNPs nearby, echoing the involvement of the pathway in JPS and HMPS. The *SH2B3* locus was associated with the combined trait of endometrial and colorectal cancer in a meta-analysis of individual GWAS (Cheng et al., 2015). Lynch syndrome provides a drive to both types of cancer; it seems that *SH2B3* variation does too, with greatly reduced penetrance.

The sporadic and familial stories are particularly tightly interwoven when it comes to *GREM1*. In this case the CRC risk association at the locus was found while searching for the cause of HMPS (Jaeger et al., 2008). The hypothesis at the time was that, as the disease was thought to be exclusive to Ashkenazim, some aspect of their shared genetic background was amplifying the effect of a common risk SNP at 15q13.3. Jaeger et al. (2008) tested 145 regional SNPs for association with sporadic CRC. Incorporating

validation sets, approximately 15,000 case-controls were genotyped, and two genome-wide significant associations with CRC obligingly appeared. The two hits were non-independent, with the stronger signal at rs4779584 ($P = 4.44 \times 10^{-14}$). rs4779584 lies within the HMPS duplication found four years later. However, all HMPS sufferers carry the protective allele in *cis* at this locus: their pathology is not directly related. Nonetheless, different changes near the same gene had been shown to underlie both syndromic and sporadic CRC predisposition.

1.4.4 Criticisms and limitations

Two initial criticisms of GWAS have become less relevant with time. The first is that controls and cases must be exquisitely matched to avoid accidentally associating genetic variation with some hidden features more common in one or other cohort. What this has ended up meaning is that GWAS must correct for population structure, or risk spurious results.

At first this was done by dividing test statistics by a constant genomic control factor, often called the inflation factor, which can be estimated from the median χ^2 statistic of tested variants. Genomic control was proposed by Devlin and Roeder (1999), who anticipated both GWAS and the problem of population structure, but not the intransigence of geneticists: early GWAS rarely if ever took this step, instead relying on recruitment criteria to avoid overt bias.

The assumption of a constant inflation statistic throughout the genome, though reasonable, is unnecessary. Instead, cryptic ancestry differences can be directly visualised through principal component analysis (PCA) (Price et al., 2006). This takes the eigenvectors of a sample covariance matrix, reducing the dimensions of the data while maximally retaining inter-sample variability. These top eigenvectors become the axes of continuous genetic variation, and the projection of the data onto these axes reveals population substructure. One of the most beautiful examples of this technique is the recreation of the map of Europe created by plotting 3000 European samples along their first two principal components (figure 1.4, Novembre et al., 2008). PCA can therefore be used both to highlight and to remove population outliers, and also to provide genotype

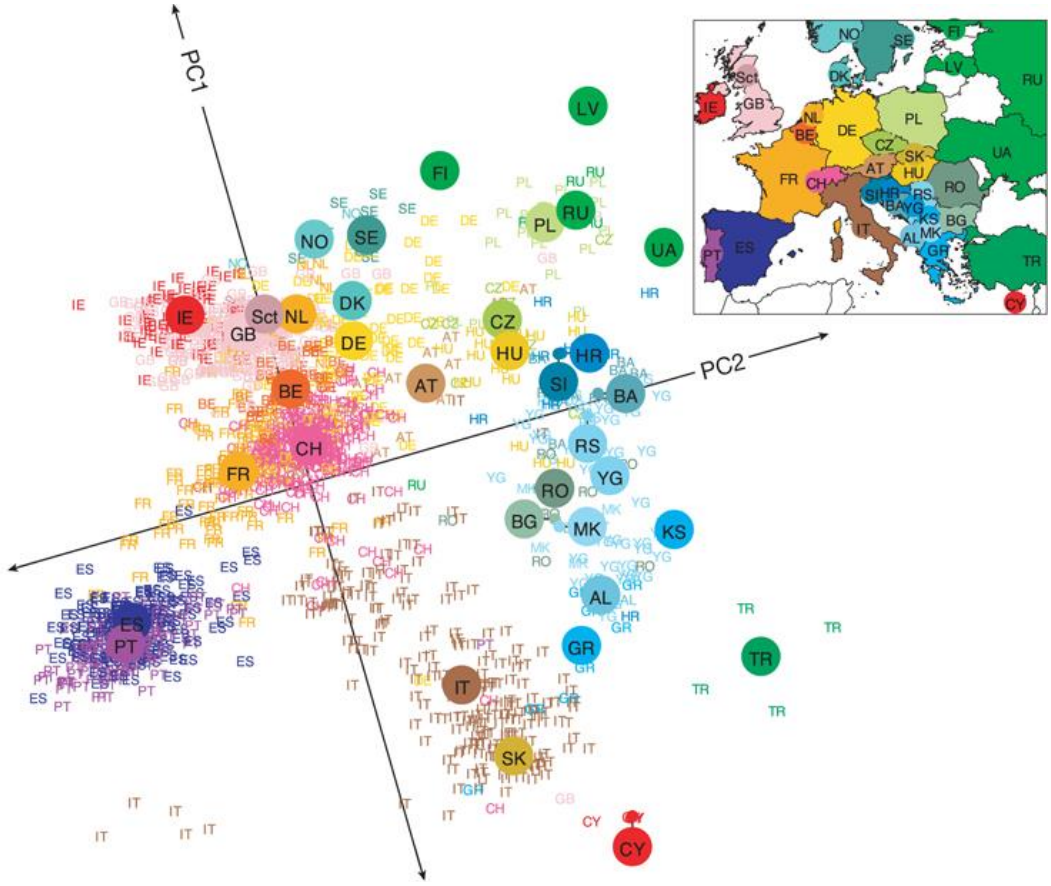


Figure 1.4: An illustration of the power of PCA from Novembre et al. (2008), showing the top two eigenvectors separating samples by geography. Colours correspond to country of origin as shown in the map, top right.

and phenotype adjustment, correcting for remaining population structure when testing for disease association. As GWAS cohorts have grown, this correction has become more important and more widespread.

This leads to the second initial criticism of GWAS: that its relatively small sample sizes were underpowered to detect association. While probably true to begin with, the advent of imputation, cheapening array technology, and international consortia have declawed this criticism for many traits. However, as I myself contend with in chapter 3, power to detect association depends on the anticipated effect size of a variant, and on its frequency. It's therefore easy to beg the question with this sort of critique, by expecting to pick up risk-associated variants which are slightly rarer and less penetrant than current cohort sizes can allow for.

We are circling a larger issue, which is the implicit disease model of GWAS. This remains a point of contention. Most studies pay a token allegiance to the common-disease-common-variants model of complex disease (CDCV). This is strongly related to Fisher's model of complex traits: a spectrum of population polymorphisms combining additively to produce a quantitative biological outcome, which could be CRC liability. As a statement about population genetics, this idea is necessarily bound up with assumptions about selection. I discuss the validity of the CDCV model in chapter 7, and find that it remains the best descriptor of complex trait genetics. However, a number of frustrations about GWAS have led some to challenge the CDCV hypothesis.

These frustrations include the patchy validation of GWAS hits, especially across populations; the lack of clear clinical utility for published results; the difficulty of attaching biological meaning to hits; and, especially, the fact that after a decade of work and huge financial investment, GWAS results only account for a minor proportion of phenotypic variance. There is missing heritability.

Many of these criticisms are unfair, and many of the suggested hiding places for the "missing heritability" are wide of the mark. As it represents a central charge against GWAS, including work in this thesis, I will discuss the missing heritability problem at length in chapter 7. But, accurate or not, the idea that GWAS and common variants have failed to explain CRC and other diseases is often used to justify a search for rarer

causative variants of intermediate effect size (Chubb et al., 2016; Gala et al., 2014; Venkatachalam et al., 2011).

1.5 Exomes, WGS, and quasi-sporadic CRC

Rarer genetic causes of CRC cannot readily be found by GWAS arrays. By happy coincidence, new technology emerged as the field was souring on GWAS. Many groups have begun to search for pathogenic coding changes in whole exome sequences of patients with early-onset CRC. These people are more likely to have a strong genetic component of disease.

At first, exome studies were more descriptive than associative. The lack of control exomes meant that groups could only highlight likely candidate variants. Smith et al. (2013) exome sequenced 50 sporadic CRC patients and found mutation and loss of heterozygosity (LoH) in *FANCM*, *LAMB4*, *LAMC3*, *PTCHD3*, and *TREX2*. Gylfe et al. (2013) similarly took 96 CRC patients with at least one affected first-degree relative, and looked for truncating mutations that also underwent LoH in tumours. *UACA*, *PSPH*, *ZNF490*, and *TSWG1* were identified as candidates. The defective *TWSG1* always exhibited LoH and was undetectable in 586 Sanger-sequenced controls; like *GREM1*, *TWSG1* is a BMP antagonist (Scott et al., 2001).

Gradually, databases of control exomes have emerged, just as in GWAS. The largest is the Exome Aggregation Consortium (ExAC, Lek et al., 2015), which has collected over 60,000 human exomes and 7.4 million variant calls, by joint-calling many smaller sets of exomes, including 1000G and the NHLBI GO Exome Sequencing Project (ESP6500). Many laboratories now also have in-house control variant frequencies, allowing a more GWAS-style comparison between rare variants.

An intermediate study mixed exome sequencing with GWAS case-control data and discovered five new candidate-wide (but not genome-wide) significant CRC-associated coding SNPs (Timofeeva et al., 2015). Oddly however, these SNPs, though rarer than typical GWAS hits, were of a similar effect size, with odds ratios below 2. Most other attempts to see strict association with one variant, hoping for a larger effect on risk, have failed. Chubb et al. (2016) compared over 1000 case and control exomes, finding no

significant association for individual variants, but an enrichment of damaging variants in cases in *IL12RB1*, *LIMK2*, and *POLE2*, and in a "DNA replication" gene ontology set (which also includes *POLE2*). Collapsing variants into gene blocks before comparing case and control frequencies has also highlighted *BUB1* and *BUB3* (de Voer et al., 2013), *BLM* (de Voer et al., 2015), *FLCN* (Dobbins et al., 2016), and *EMR3*, *PTPN12*, and *LRP6* (de Voer et al., 2016). This is an association strategy I discuss more in chapter 7.

These results need to be replicated, though this may be by whole-genome sequencing (WGS), the newest addition to the increasingly expensive genetics toolbox. It's also hoped that WGS will discover non-coding predisposition variants. The perennial problem of control cohorts has begun to be solved, not least by the WGS500 project (Taylor et al., 2015), which is a collection of 500 whole-genome sequences of people with a wide spread of pathological phenotypes, who can to an extent control for each other.

I haven't seen it put in these terms, but the idea that some CRC predisposition is caused by a few, or even singular, rare variants with moderate penetrance is a mild rejection of the Mendelian/sporadic dichotomy constructed at the start of this thesis. CRC cases caused by these variants would be "quasi-sporadic", displaying neither the monogenic high penetrance of Mendelian disease nor the additive curve of polygenic input, subordinate in the most part to chance, that is assumed to characterise sporadic CRC. This potential rarefaction of accounts of CRC heredity recalls previous category refinements: of cancer in general into tissue types; of heritable CRC into named disorders with different causes; of CRC into CMS groups. Perhaps ever-forking subsets are an inevitable outcome of biology.

1.6 Outline of Thesis

The first half of this thesis is concerned with sporadic CRC. In chapter 3, I impute a denser genotype array at the *GREM1* locus, attempting to refine our model of local CRC risk. I then explore the biological function of the novel association uncovered with a mixture of bioinformatic and molecular assays. Chapter 4 performs simliar fine-mapping at the *POLD3* locus. I also examine to what extent GWAS SNPs overall impinge on

CRC incidence, outcome, and mutation burden.

The second half of the thesis is an exploration of a set of nearly 500 UK genomes. I describe the remapping and calling of this cohort in chapter 5, and its carriage of mutation within genes known to cause Mendelian disease, especially Mendelian CRC syndromes. Moreover, I ask to what extent population structure is visible in a disordered, clinically recruited set of this type. Then, in chapter 6, I use a variety of filtering strategies to search these genomes and additional whole-exome sequences for pathogenic germline variation, looking in particular detail at MMR genes and at individuals with available family genotypes.

Chapter 7 puts the findings of the thesis into context. In particular, I discuss the problem of "missing heritability", and the extent to which whole-genome sequencing is an improvement on GWAS.

Chapter 2

Materials and methods

2.1 Ethics

All work in this thesis was performed with UK national ethical committee approval, including patient recruitment, anonymised tissue collection, and tumour molecular analysis. Individuals in this thesis had provided written informed consent for the undertaking of genetic investigation into the cause of their disease.

In particular, CORGI recruitment and genotyping was covered by ethics council MREC/06/Q1702/99; "Carlos" exome sequencing and VICTOR/QUASAR2 analyses were approved by ORECB/05/Q1605/66; and "Cheadle" exome sequencing was approved by REC/04/MRE06/60.

2.2 GWAS cohorts

For the fine-mapping of colorectal cancer risk association in chapters 3 and 4, a number of case-control datasets were used. These are listed below and referred to consistently in the thesis by the acronyms given. The cohorts are principal component analysis (PCA)-cleaned (see §2.7.8) and free of duplicates, related individuals, and other classification errors. As such the numbers reported may not match those in the earliest publications using each dataset; an accurate report of why samples were removed is given in supplementary figure 1 of Dunlop et al. (2012), and supplementary figure 3 of Al-Tassan et al. (2015), whose inclusion criteria I share.

The inclusion in all GWAS cohorts of patients with a family history of CRC and

relatively early onset is both a strength and a weakness. In theory, incorporating individuals with an apparently greater genetic disease component enriches for higher effect size variants. However, this bias also means the precise phenotypic category in question is blurred. The genetic architecture of "truly" sporadic CRC could differ significantly from that of quasi-Mendelian cancer.

2.2.1 LP1

London Phase 1 (LP1; Tomlinson et al., 2008) drew 888 cases from the *Co*lo*R*ectal Tumour *G*ene *I*dentification consortium (CORGI). These 888 patients had been diagnosed with CRC before the age of 75, and most also had family history of CRC. English hospitals and genetics clinics recruited the cases, and known polyposis syndromes, Lynch syndrome, and bi-allelic *MUTYH* mutations were screened out. All CORGI recruitments in other data sets share these criteria.

LP1 controls were 998 partners of the recruited cases who were themselves unaffected by colorectal cancer, lacking a family history of CRC to second-degree relative level. All LP1 samples were genotyped using the Illumina Hap550 SNP array.

2.2.2 LP2

London Phase 2 (LP2; Tomlinson et al., 2011) cases were recruited through two initiatives: the National Study of Colorectal Cancer Genetics, and the Royal Marsden Hospital Trust/Institute of Cancer Research Family History and DNA Registry. These collected 2659 CRC cases (mean age at diagnosis 59.3 years) and 2798 controls (mean age 59.8 years). All recruited participants had self-reported European ancestry. Controls were the unrelated friends or spouses of the patients without history of CRC. Genotyping was performed on Illumina iSelect and Illumina Goldengate.

2.2.3 SP1

Scotland Phase 1 (SP1; Tenesa et al., 2008) included 973 CRC cases aged less than 55 at diagnosis. Patients with known polyposis syndromes, Lynch syndrome, or bi-allelic *MUTYH* mutation were excluded from the study. 998 SP1 controls were sampled

from Scottish NHS registers, matched by age (±5 years), gender, and area of residence within Scotland. Genotyping was performed on Illumina HumanHap300 and Illumina HumanHap240S arrays.

2.2.4 SP2

Scotland Phase 2 (SP2; Houlston et al., 2010) comprised 2007 CRC cases (mean age at diagnosis 65.8 years) and 2075 population controls (mean age 67.9 years) from Scotland. Cases were taken from a prospective incident CRC case series and were aged less than 80 years at diagnosis. The population-based control subjects were frequency matched by age, gender, and area of residence within Scotland. Genotyping was performed on Illumina iSelect and Illumina Goldengate.

2.2.5 CFR1

The Colon Cancer Family Registry 1 (CFR1; Newcomb et al., 2007) set comprised 1175 American, Canadian, and Australian familial CRC cases and 999 controls (coloncfr.org). The cases were recruited by the Seattle Familial Colorectal Cancer Registry, the Ontario Familial Cancer Registry, and the Australasian Colorectal Cancer Family Study. Controls were from the same populations; all were of non-Hispanic white ethnicity. CFR1 genotyping was performed on Illumina Hap550.

2.2.6 CFR2

The CFR2 cohort was a further 795 cases from the Colon Cancer Family Registry, recruited from centres in Australia, Ontario, Seattle, USC, Mayo, and Hawaii, and 2242 controls from the Cancer Genetic Markers of Susceptibility (CGEMS) studies of breast ($n = 1141$) and prostate ($n = 1101$).

The 1141 CGEMS breast controls were from the Nurses' Health Study, a longitudinal study of cancer-free women who provided blood samples between 1989–1990 (Hunter et al., 2007). The women were followed up for incidental disease until 2004, and genotyped on Illumina Hap550.

CGEMS prostate controls were from the Prostate, Lung, Colon, and Ovarian Cancer

Screening Trial (Yeager et al., 2007). This trial recruited 155,000 people aged 55–74 with no history of these cancers; 1101 men with no history of prostate cancer were genotyped on Illumina Hap550 and used as CFR2 controls.

2.2.7 VQ58

The VQ58 GWAS set is a combination of two patient trials and separate controls. The cases were genotyped on a combination of Illumina HumanHap300, Illumina HumanHap270, and Illumina Human 1.2MDuo Custom_V1arrays. Controls were genotyped on Illumina 1.2M arrays as part of the Wellcome Trust Case Control Consortium (2007). Due to VQ58's use in both chapter 3 and, along with clinical data from the trials, in chapter 4, I have explained its structure in more detail below.

2.2.7.1 VICTOR

*Vi*oxx in *C*olorectal Cancer *T*herapy: Definition of *O*ptimal *R*egime (VICTOR; oncology.ox.ac.uk/trial/victor) was a randomised controlled trial testing the selective COX-2 inhibitor rofecoxib (Vioxx) in CRC patients after therapy, to assess whether the drug could reduce cancer recurrence (Midgley et al., 2010; Pendlebury et al., 2003).

Vioxx was withdrawn by Merck in 2004 due to its unforeseen effects on cardiovascular risk. The drug increases the risk of heart attack and stroke, an outcome first visible in the VIGOR study (Bombardier et al., 2000), which was part-authored by Merck employees. The alleged misreporting of risk outcomes in Bombardier et al. (2000) led to a fierce exchange between authors and *New England Journal of Medicine* editors (Bombardier et al., 2006; Curfman et al., 2005) which spread to the courts (Plunkett v. Merck, 2005). However, Merck voluntarily withdrew the drug after again seeing increased risk of cardiovascular pathology in their APPROVe study (Bresalier et al., 2005). This was two years into VICTOR's recruitment phase; the trial was immediately abandoned. Even with a shortened treatment time, adverse cardiovascular events occurred more frequently in the rofecoxib arm of VICTOR than would be expected by chance (Kerr et al., 2007).

VICTOR supplied 921 CRC cases to the VQ58 set.

2.2.7.2 QUASAR2

*Qui*ck *A*nd *S*imple *A*nd *R*eliable 2 (QUASAR2; oncology.ox.ac.uk/trial/quasar-2) was a multicentre study comparing stage II and III CRC treatments: capecitabine alone, or capecitabine supplemented with bevacizumab, a VEGF-A inhibitor.

QUASAR2 supplied 873 cases to VQ58.

2.2.8 BC58

The 2686 VQ58 controls were from two sources: the 1958 Birth Cohort (BC58), a subset of the National Child Development Study (Power and Elliott, 2006), and the National Blood Service, a UK blood transfusion services repository of anonymised DNA samples. The BC58 was an epidemiological survey based on all individuals born in England, Wales, and Scotland during one week in 1958 (www.b58cgene.sgul.ac.uk).

2.3 Validation GWAS Cohorts

2.3.1 COIN

The COIN GWAS set (Al-Tassan et al., 2015) was based on 2244 CRC cases (mean age at diagnosis 61 years) recruited to two Medical Research Council trials designed to assess the benefits of *CO*ntinuous versus *IN*termittent chemotherapy in advanced CRC: COIN and COIN-B (Maughan et al., 2011; Wasan et al., 2014). COIN patients were randomised into three groups: one received continuous oxaliplatin and fluoropyrimidine chemotherapy, one continuous chemotherapy plus cetuximab, and one intermittent chemotherapy. COIN-B patients all received intermittent chemotherapy but were randomised between continuous and intermittent cetuximab supplement. The COIN and COIN-B trial cases were genotyped on Affymetrix Axiom arrays.

2162 COIN GWAS controls were members of the WTCCC2 database, as in VQ58. They were genotyped on the Affymetrix 6.0 array.

2.3.2 Finland

The Finnish GWAS cohort is 1172 CRC cases and 8266 cancer-free controls ascertained through Finnish Hospitals and the Finnish Cancer Registry. Cases were genotyped using Illumina HumanOmni 2.5M8v1; controls were genotyped on a combination of Illumina HumanHap 670k and 610k arrays. Both were imputed on a combined reference panel of the 1000G phase 1v3 and 3882 population-matched Sequencing Initiative Suomi haplotypes (Orlando et al., 2016).

2.3.3 CORGIbcd

An extension of LP1, this validation cohort consisted of 562 cases and 910 controls recruited though the CORGI study and sharing its inclusion criteria.

2.3.4 ICR

Dr. Peter Broderick of the Institute of Cancer Research (ICR) genotyped rs17816465 in 3227 CRC cases and 3613 controls.

2.3.5 CHIBCHA

The genetic study of *C*ommon *H*ered*i*tary *B*owel *C*ancers in *H*ispania and the *A*mericas (CHIBCHA) is a growing collection of GWAS data looking for novel CRC associations in admixed Latin American samples. I was kindly provided by Dr. Luis Carvajal-Carmona with PCA-corrected summary statistics for rs17816465 association in two CHIBCHA cohorts: "Colombia" (820 cases, 822 controls) and "Mexico" (904 cases, 866 controls).

2.3.6 CORECT

The CORECT consortium collected six smaller CRC GWAS sets, one of which is a larger version of CFR2 (§2.2.6). The six are:

1. MECC2 (1120 cases, 820 controls), from Northern Israel
2. CFR2 (1660 cases, 1393 controls), from America and Australia
3. Kentucky (1038 cases, 1134 controls), from America

Study	Cases	Controls	Chip	Reference
ARCTIC	601	519	see below	Cotterchio et al. (2005)
ASTERISK	892	947	CytoSNP	Küry et al. (2007)
COLO 2&3	87	125	CytoSNP	Le Marchand et al. (2001)
DACHS Set 1	1710	1708	CytoSNP	Brenner et al. (2011)
DACHS Set 2	666	498	CytoSNP	Vossen et al. (2011)
DALS Set 1	706	710	550/610	Slattery et al. (1996)
DALS Set 2	410	464	CytoSNP	Slattery et al. (1996)
HPFS Set 1	227	230	OmniExpress	Rimm et al. (1990)
HPFS Set 2	176	172	OmniExpress	Rimm et al. (1990)
HPFS AD	313	345	OmniExpress	Rimm et al. (1990)
MECC2	328	346	CytoSNP	Kolonel et al. (2000)
NHS Set 1	391	774	OmniExpress	Belanger et al. (1978)
NHS Set 2	158	181	OmniExpress	Belanger et al. (1978)
NHS AD	513	578	OmniExpress	Belanger et al. (1978)
PHS Set 1+2	375	389	OmniExpress	Christen et al. (2000)
PLCO Set 1	533	1976	550/610	Gohagan et al. (2000)
PLCO Set 2	486	415	CytoSNP	Gohagan et al. (2000)
PMH-CCFR	280	122	CytoSNP	
VITAL	285	288	CytoSNP	White et al. (2004)
WHI Set 1+Hip	313	1467	550/610	Hays et al. (2003)
WHI Set 2	651	816	CytoSNP	Hays et al. (2003)
WHI WGS	610	309	-	Hays et al. (2003)
Total	10,711	13,379		

Table 2.1: The constituent cohorts of GECCO. WHI WGS is 6× WGS data, against which the other cohorts were imputed using MACH. ARCTIC was genotyped on a mixture of the Affymetrix GeneChip Human mapping 100K and 500K Array Set and a 10K nonsynonymous SNP chip; CytoSNP, Illumina CytoSNP BeadChip; 550/610, Illumina 610K and 550K.

4. CPS-II (548 cases, 538 controls), from America
5. NFCCR (195 cases, 477 controls), from Newfoundland
6. Melbourne (539 cases, 469 controls), from Australia, UK, Italy, and Greece

All were genotyped on a custom Affymetrix genome-wide platform, the Axiom CORECT Set, with $\sim$ 1.3 million SNPs and indels on two chips. CORECT genotypes were subsequently enriched by imputation on 1000G. The membership total of CORECT was therefore 5100 cases and 4831 controls (Schumacher et al., 2015).

2.3.7 GECCO

The *G*enetics and *E*pidemiology of *C*olorectal Cancer *CO*nsortium (GECCO) coordinating centre has collected 22 GWAS case-control cohorts, and allowed me access to summary data from a meta-analysis thereof. The constituent cohorts are summarised in table 2.1; for more detail see Hiraki et al. (2013).

2.4 WGS cohorts

Individuals whose genomes were sequenced were, like those genotyped for GWAS, often bearers of obviously familial, severe, or early-onset disease. This bias is less problematic in the context of WGS. The genomes were used without assuming a single disease model and were not considered a homogenous collection.

2.4.1 Illumina 215

The 215 individuals within the "Illumina 215" were ascertained through CORGI, and were required to have either over 5 adenomas or CRC before age 65, or to have a first-degree relative with these characteristics.

2.4.2 WGS500

As part of the Oxford-Illumina WGS500 project, 65 genomes of CRC patients were whole-genome sequenced (Taylor et al., 2015). 15 such genomes were described in Palles et al. (2013), which allowed the discovery of polymerase proofreading-associated polyposis.

2.4.3 HPPS

The "HPPS" set consisted of 24 whole genomes and 52 exomes. The aim of this cohort was to sequence individuals who fulfilled WHO criteria for serrated polyposis syndrome (SPS; Snover et al., 2010) in number of polyps if not in location or size, as this information was often lacking. See §6.1.2. 6 genomes and 25 exomes had ≥ 5 serrated polyps and fulfilled criterion *i.*; 13 genomes and 15 exomes fulfilled criterion *iii.* in having ≥ 20 polyps. The remaining 5 genomes and 12 exomes showed some evidence of serrated polyposis in their family history.

2.4.4 CGI196

The Complete Genomics CRC cases (CGI196) were ascertained through CORGI. They all had at least one first-degree relative with CRC, and had been diagnosed before

age 55 (Whiffin et al., 2013). The genomes were sequenced and called with Complete Genomics-exclusive software (Drmanac et al., 2009).

2.5 Exome sequences

Variant filtering in §6.2.1 made use of five collections of exome sequencing, listed in table 2.2. These exomes were supplied as different levels of summary data, and could not therefore be called together.

Name	Samples	Reference
Carlos	20	López-García et al. (2017)
Cheadle	42	Maughan et al. (2011); Smith et al. (2013)
HPPS	52	see §2.4.3
Mseq	17	Cross et al., manuscript in preparation
Oliver	107	Sieber et al. (2003)

Table 2.2: Exome sequences

2.6 Bioinformatics

2.6.1 WGS pipeline

All genotyping and quality control steps were performed according to GATK best practices at the time (Van der Auwera et al., 2013).

2.6.1.1 Remapping

The Illumina 215 genomes were originally mapped with the Isaac Aligner (Raczy et al., 2013). The HPPS genomes had been mapped with BWA/Stampy (Li and Durbin, 2009; Lunter and Goodson, 2011), but to a reference genome distinct from the WGS500 alignments (hg19 rather than GRCh37d5). Therefore, both sets required remapping.

To achieve this, Dr Palles and I made use of an in-house pipeline written by Dr Francesc Castro. This invokes a combination of SAMtools, BEDtools, fastqutils, BWA/Stampy, and the Picard suite to generate split fastq files from mapped BAMs, and then to realign the output and deduplicate the new BAM files (see §2.6.2 for software details).

2.6.1.2 BAM recalibration

On a per-chromosome and per-individual basis, I processed remapped BAMs with GATK IndelRealigner using known sites from the Mills and 1000 Genomes (1000G) gold standard indels VCF and 1000G phase 1 indels VCF, after realigner target creation. (This step is now deprecated, as HaplotypeCaller no longer requires indel realignment prior to use). Base quality score recalibration followed, with GATK's BaseRecalibrator constructing the covariation model while masking known polymorphisms present in either of the two files above or in dbSNP 138. GATK PrintReads then adjusted the base quality scores according to this model.

2.6.1.3 Genotyping

I used HaplotypeCaller to create gVCF-format files (which retain invariant reference calls) with the DISCOVERY genotyping mode, a minimum phred-scaled Q score confidence threshold of 10 for emission and of 30 for high-confidence calling.

2.6.1.4 Post-genotyping quality control

I combined gVCFs per-chromosome into multisample files using CombineGVCFs. I then performed GATK-recommended quality control to recalibrate variant scores using training sets: once for SNPs, once for indels. The training sets were provided in the GATK resource bundle. For SNPs, true sites used as training resource were the Hapmap and Omni variants (priors of Q15 and Q12 respectively); "false" (false positive-containing) sites used in training were the 1000G phase 1 high-confidence SNPs (prior of Q10); and dbSNP 138 sites were considered false and not used in training, but were still used to stratify variant metrics (prior of Q2). The annotations DP, QD, FS, SOR, MQ, MQRankSum, ReadPosRankSum, and InbreedingCoeff were used for calculations. For indels, the Mills and 1000G gold standard indels were considered true with a prior of Q12, and were used in training. GATK's ApplyRecalibration tool used the error model defined by output recalibration and tranches files to perform VQSR in each case, keeping tranche 99.9% and above.

I then decomposed the recalibrated VCFs into purely biallelic lists using vt in smart

Description	Source	Date/version
1000G alternative allele frequency	www.1000genomes.org	02/05/20
ExAC alternative allele frequency	exac.broadinstitute.org	0.3
ESP6500 variant frequency	evs.gs.washington.edu/EVS	01/07/2014
UK10K variant frequency	www.uk10k.org	22/07/2014
COSMIC accession number	cancer.sanger.ac.uk/cancergenome/projects/cosmic	10/02/2014
Super duplication	genome.ucsc.edu/cgi-bin/hgTables	17/11/2014
Phylo-P score	genome.ucsc.edu/cgi-bin/hgTables	26/06/2013
Phastcons score	genome.ucsc.edu/cgi-bin/hgTables	26/06/2013
REPMASK class	genome.ucsc.edu/cgi-bin/hgTables	27/11/2014
HGMD PRO data	www.biobase-international.com/product/hgmd	01/04/2014
GERP score	mendel.stanford.edu/sidowlab/downloads/gerp	01/05/2011
ClinVar data	www.ncbi.nlm.nih.gov/clinvar	03/02/2015
WGS500 variant frequency	freeze5	27/12/2014
CGI96 variant frequency	-	25/06/2016
SYSCOL ChIP-seq peaks	walnut.biosci.ki.se	24/02/2016

Table 2.3: Annotations on the Tomlinson genomes. Top, public databases; bottom, annotations from in-house databases. All UCSC-derived annotations are for human genome build 37.

mode, prior to further annotation. The one filter always applied was a maximum missingness of 25% per variant using VCFtools; further filtering and subsetting was project-dependent.

2.6.1.5 Annotation

Annotation of the Tomlinson genomes was performed both by myself and Dr. Silvia Salatino. The main tool was the Ensembl VEP (Variant Effector Predictor) framework (table 2.5); nonstandard annotations and databases used are listed in table 2.3. The SYSCOL ChIP-seq peaks were performed by members of the Taipale laboratory in the colorectal cell line LoVo, and were pulldowns of the transcription factors CTCF, GATA2, GATA4, GATA6, MAX, MYC, RAD21, SMAD1, and TCF7L2.

2.6.2 Software

Software used for GWAS and array work is listed in table 2.4. Software used for WGS work is listed in table 2.5. Other generic software used in this thesis includes the academic java applet IBS v1.02 (Liu et al., 2015a), used in figure 6.8c; and IGV v2.3.91 (Robinson et al., 2011), used in figures 6.3b and 6.4d.

The statstical software R version 3.2.2 was used throughout (R Core Team, 2012). Packages of note used in thesis include `LDheatmap` (Shin et al., 2006), `dplyr` (Wickham

Name	Version	Use	Citation
PLINK	1.07	genotype file management, LD	Purcell et al. (2007)
	1.90b3.29	measurements, haplotype testing	Chang et al. (2015)
SHAPEIT	2.r837	phasing	Delaneau et al. (2012)
IMPUTE2	2.3.0	genotype imputation	Howie et al. (2012); Marchini et al. (2007)
GTOOL	0.7.5	PLINK to SNPTEST file conversion	www.well.ox.ac.uk/ cfreeman/software/gwas/gtool.html
SNPTEST	2.4.1	association testing	used in Wellcome Trust Case Control Consortium (2007)
META	1.4	meta-analysis of associations	Liu et al. (2010)
LocusZoom	1.3	visualising associations	Pruim et al. (2010)
SNAP	2.2	LD and proxy information	Johnson et al. (2008)
EIGENSOFT	6.0.1	PCA calculation and correction	Price et al. (2006)
PEDSTATS	0.6.10	MERLIN file creation	Wigginton and Abecasis (2005)
MERLIN	1.1.2	linkage testing	Abecasis et al. (2002); Abecasis and Wigginton (2005)

Table 2.4: Sotware used with GWAS genetic data, whether from array or extracted from sequencing.

Name	Version	Use	Citation
fastqutils	within ngsutils 0.5.7-14d103a	converting BAM to and from FASTQ	Breese and Liu (2013)
SAMtools	1.2	indexing and writing BAM files	Li et al. (2009)
Picard	2.1.1	BAM deduplication	broadinstitute.github.io/picard
BWA	0.7.10-r1005-dirty	genome alignment	Li and Durbin (2009)
Stampy	1.0.2	mapping	Lunter and Goodson (2011)
Platypus	0.7.9.3	calling	Rimmer et al. (2014)
GATK	3.5-0	calling and filtering	McKenna et al. (2010)
vt	v0.5772-d00d603a	decomposing variants	Tan et al. (2015)
VEP	86	VCF annotation	McLaren et al. (2016)
Pindel	0.2.4w	calling large structural changes	Ye et al. (2009)
HTSlib	1.3.2-134-g1bc5c56	compressing and indexing VCF files	www.htslib.org
BCFtools	1.3.1-163-gbbd1e4f	VCF manipulation	www.htslib.org
BEDOPS	2.4.19	BED file manipulation	Neph et al. (2012)
BEDtools	2.25.0	BED file manipulation	Quinlan and Hall (2010)
LiftOver	-	genome coordinate conversion	hgdownload.cse.ucsc.edu

Table 2.5: WGS tools used in this thesis

and Francois, 2015), `RColorbrewer`, `survival` (Therneau, 2015), `Gviz` (Hahne et al., 2014), `genetics` (Warnes et al., 2013), and `hapassoc` (Burkett et al., 2006).

2.6.3 Imputation

Missing genotypes were imputed using IMPUTE2 (Howie et al., 2012, , see table 2.4). Effective population size was set at 20,000, as suggested; ten "burn-in" Monte Carlo Markov Chain iterations were followed by twenty iterations that contributed to final imputation probabilities (the default settings). IMPUTE2 output was then filtered on MAF ($> 1\%$ in cases and controls), missing data proportion (≤ 0.05), and info score (≥ 0.95).

For fine-mapping, I used a combined reference panel. Haplotypes from phase 1v3 of 1000G were combined with phased high-coverage ($> 30\times$) whole-genome sequences of 196 British individuals (CGI196; see §2.4.4). This addition increases imputation accuracy (Timpson et al., 2014).

2.6.4 Association testing

The established test for association of genotype and phenotype is Pearson's χ^2 test on a 2×3 contingency table (which has two degrees of freedom; see §2.7.1). The rows on this table are case/control status; the columns are the three possible genotypes. The table can shrink to 2×2 under different models, for example if one wishes to test a dominant (AA vs Aa and aa) genotypic effect. However, this thesis only considers additive models, a standard assumption in most GWAS and post-GWAS work.

2.6.5 CpG island definition

The CpG islands at *GREM1* and *KCTD10* were defined using the UCSC's own criteria, which are:

- GC content of 50% or greater
- Length greater than 200bp
- Ratio greater than 0.6 of observed number of CG dinucleotides to the expected number on the basis of the number of Gs and Cs in the segment, as defined by

Gardiner-Garden and Frommer (1987):

$$\text{Obs/Exp CpG} = \frac{\text{Number of CpG}}{\text{Number of C} \times \text{Number of G}} \times N \tag{2.1}$$

where N is the length in nucleotides of the sequence in question.

2.6.6 Duplication mapping from BAMs

I looked for large-scale structural changes in WGS data in two ways. First, by down-sampling VCF files to Illumina HumanOmniExpress-24 v1.0 BeadChip variants, and converting this quasi-array data to B-allele frequency data using a modified script by Dr. Sam Knight, which could then be examined using Illumina's Nexus Copy Number software (v7.0). Second, by direct calling with Pindel v0.2.4w, which takes BAM file input.

2.6.7 ChIP-seq analysis

ChIP-seq data were analysed by Dr. Rafik Salama of the Ratcliffe laboratory following their in-house pipeline. Briefly, this consisted of FastQC adaptor trimming and read quality control (www.bioinformatics.babraham.ac.uk/projects/fastqc), alignment to hg19 with BWA (Li and Durbin, 2009), sorting and duplicate marking with the Picard tools markDuplicates and sortSAM (see table 2.5), peak calling with T-PIC (Hower et al., 2011), and motif recognition by MEME-ChIP v4.9.0 (Machanick and Bailey, 2011). Called peaks were analysed with the webtool GREAT v3.0.0 (McLean et al., 2010) for nearby gene enrichment by ontology.

2.6.8 Linkage analysis

For the two family linkage analyses in chapter 6, I extracted genotypes from a number of sources. The genetic data of the four members of OX2027 consisted of one whole genome sequence, one member of CORGI (§2.2.1) genotyped on Illumina Hap550, and two individuals genotyped on the Illumina HumanCoreExome chip. The five OX70783 cohort members divided into two genotyped by WGS, one genotyped on the Illumina

HumanCoreExome chip, and two genotyped on the Illumina Axiom Biobank genotyping array.

Dr. Claire Palles had phased and imputed the three chip datasets to a reference of 1000G Phase 3 (October 2014). I therefore extracted a subset of SNPs from each using a combination of GTOOL, PLINK, and BCFtools (§2.6.2). This subset was first defined in CORGI as polymorphisms in Hardy-Weinberg equilibrium ($P < 0.000001$), with per-sample missingness below 2%, per-variant missingness below 5%, $r^2 < 0.4$ with other SNPs, and imputation info score > 0.8. The intersection of this SNP list with SNPs also imputing with info> 0.8 from the other genotyping platforms defined the variants used to map linkage: 170,362 for OX2027, and 132,124 for OX70783. I used MERLIN to perform a nonparametric, centimorgan-binned test of Kong and Cox *LOD* score (Kong and Cox, 1997). MERLIN outputs both linear and exponential K&C *LOD* scores; the exponential model was more appropriate in this thesis, where single pedigrees only were being tested for a large increase in allele sharing, and is the statistic on which I focus.

2.7 Statistical methods

2.7.1 χ^2 tests

Pearson's χ^2 test is used to describe the strength of association bewteen data categories in contingency tables. It generates the χ^2 statistic

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{2.2}$$

where n = the number of cells in the table, O_i = the the number of observations of type i, and E_i = the expected frequency of type i.

The null distribution of a Pearson statistic with j rows and k columns is approximated by the chi-squared distribution with $(k-1)(j-1)$ degrees of freedom, an approximation which improves with increasing sample size. Integration to the right of a generated χ^2 statistic on this null distribution provides a P value.

Pearson's test was most commonly used in this thesis to compare allele frequencies between cases and controls in GWAS, using PLINK (§2.6.2).

2.7.2 Logistic regression

Logistic regression is a generalised linear regression of a variable on the natural log (the "logit transform") of a binary or proportional outcome. In this thesis this outcome is mostly CRC disease status. The log of the odds of CRC is related to the number of alleles carried, normally in an additive model, with additional independent variables added if needed, such as age, sex, or genotype status at other loci. These regressions can be conducted iteratively as normal. I performed logistic regression in chapters 3 and 4 using a mixture of SNPTEST (Marchini and Howie, 2010) and the statistical software R.

2.7.3 Information criteria

For a given model, the Aikake information criterion AIC for that model is

$$AIC = 2k - 2ln(L) \tag{2.3}$$

where k is the number of parameters, and L is maximised likelihood function of the model. It is a formalised way of weighing goodness of fit against model complexity, which is penalised by the $2k$ term: the smaller the AIC, the qualitatively better the model. In chapter 4 this thesis additionally uses the related Bayesian information criterion, BIC

$$BIC = ln(n) \cdot k - 2ln(L) \tag{2.4}$$

where n is sample size. This is almost identical, but more punitively punishes parameter inclusion.

2.7.4 Multiple testing correction

GWAS performs very many tests on the same outcome data, vastly increasing the likelihood of false positive associations. It's standard to use the Bonferroni correction to diminish this risk: the P value threshold for significance is divided by the number of non-independent tests performed. For example, the 1,000,000 SNPs in early GWAS

means that "genome-wide significant" has come to mean $P < 5 \times 10^{-8}$. This threshold is adopted here.

Ouside of GWAS, in comparisons between multiple means this thesis uses analysis of variance (ANOVA) followed by Tukey's honest significant difference (HSD) *post hoc* test. Tukey's HSD corrected P value is the difference between pairs of means over the standard error of the data; the statistic has a Studentised range distribution. Tukey's HSD is similar to a t test, but corrects for multiple testing.

P values of expression quantitative trait loci (eQTL, §2.7.12) coefficients were corrected for multiple testing by the Benjamini-Hochberg false discovery rate method, which returns a corrected q value equal to the smallest familywise alpha (false discovery rate) under which the uncorrected test would have been rejected.

2.7.5 Power calculations

The power of a test is the probability of that test correctly rejecting the null hypothesis in the case of real signal. A traditional threshold for a "well-powered" test is 80% (de Bakker et al., 2005). Power for GWAS testing was calculated using a non-central "null" χ^2 distribution (Patnaik, 1949). This distribution's non-centrality parameter, λ, was calculated by equation 2.5.

$$\lambda = \frac{2r^2 N_{\text{case}} N_{\text{cont}} (F_{\text{case}} - F_{\text{cont}})^2}{NF(1-F)} \tag{2.5}$$

where N is sample size, F is risk allele frequency, and r^2 is the LD between marker of interest and causal variant, which I held at $r^2 = 1$. Power calculations were performed by modifying an R script by Damjan Vukcevic.

2.7.6 IBD

$$\hat{\pi} = \frac{P(Z=1)}{2} + P(Z=2) \tag{2.6}$$

Where $P(Z = 1)$ and $P(Z = 2)$ are the probabilities of two samples sharing one or two alleles at a locus, respectively. These probabilities depend on allele frequency and are averaged over all SNPs tested, to give $\hat{\pi}$, a measure of relatedness between samples

based on identity by descent (IBD) sharing. Identical twins (or duplicate samples) will have $\hat{\pi} = 1$; first-degree relatives will have $\hat{\pi} = 0.5$; second-degree relatives have $\hat{\pi} = 0.25$, and so on. IBD was calculated with PLINK 1.9 using SNPs from the Omni1M chip in chapter 5.

2.7.7 Linkage disequilibrium

Under Mendel's third law, genetic loci assort independently. An allele A_1, at biallelic locus A with frequency P_{A_1}, should be totally independent of allele B_1, present at B with frequency P_{B_1}, such that

$$D = P_{A_1B_1} - P_{A_1}P_{B_1} = 0 \tag{2.7}$$

Which is to say, the observed frequency of the A_1B_1 haplotype, $P_{A_1B_1}$, is identical to the frequency one would expect by chance. In this case the loci are in "linkage equilibrium". Independent assortment does not hold for syntenic loci, which must be separated by recombination, as discussed in §1.2.1. The magnitude of D when $D \neq 0$ therefore quantifies the interdependence of the loci; D is the coefficient of linkage disequilibrium.

As is standard in order to allow meaningful comparisons of linkage between different allele pairs, this thesis only reports scaled versions of D. The first is Lewontin's D', which divides D by its theoretical maximum D_{max}, where

$$D_{max} = \begin{cases} min\{P_{A_1}P_{B_1},\, P_{A_2}P_{B_2}\} \text{ if } D < 0 \\ min\{P_{A_1}P_{B_2},\, P_{A_2}P_{B_1}\} \text{ if } D > 0 \end{cases} \tag{2.8}$$

The second is simply the square of Pearson's correlation coefficient between the loci, r^2, where

$$r^2 = \frac{D^2}{P_{A_1}P_{A_2}P_{B_1}P_{B_2}} \tag{2.9}$$

In both cases P_{A_2} and P_{B_2} are the frequencies of A_2 and B_2, the other alleles at each locus.

D' is a better descriptor of coinheritance and is most useful when considering hap-

lotype structure. r^2, perhaps a cruder instrument, is useful when considering how well one genotype can substitute for another—for example, if assessing the descriptive potential of a proxy SNP. The measurements tend to diverge due to allele rarity, since allele coincidence as expressed by r^2 will decrease as one allele becomes rarer, even if it lies in total disequilibrium with a second locus.

2.7.8 PCA

Eigendecomposition of a normalised genotypic covariance matrix provides the orthogonal basis vectors of that space, called its principal components. These eigenvectors can be ordered by the amount of variance they explain (their eigenvalues): plotting the covariance matrix on the largest of these axes (the first few principal components) is a way of visualising its internal structure, as discussed in chapter 1 and shown in figure 1.4.

To perform PCA on the Tomlinson genomes in chapter 5, I used the EIGENSTRAT software (Price et al., 2006) on SNP subsets as defined in the text. This is part of the EIGENSOFT package; see table 2.4. Relatives were not used to calculate components, but were projected onto the axes afterwards.

2.7.9 Clines

To establish rare variant sharing enrichment along geographic clines, I created a distance matrix of the 238 individuals with satisfactory address estimates who were neither related nor obviously of non-British ancestry. I did this by constructing a simple xml script to convert postcodes into longitude and latitude from the Google maps database (maps.google.com/maps/api/geocode). I then calculated the Euclidean distance between each position using the Haversine formula:

$$x = \sin(\frac{\frac{\pi}{180}lat_2 - \frac{\pi}{180}lat_1}{2})^2 + \cos(\frac{\pi}{180}lat_1).\cos(\frac{\pi}{180}lat_2).\sin(\frac{\frac{\pi}{180}long_2 - \frac{\pi}{180}long_1}{2})^2 \tag{2.10}$$

$$d = 2r \times atan2(\sqrt{x}, \sqrt{1-x}) \tag{2.11}$$

Where r is the radius of the Earth (using $r = 6378.145$km), d is the calculated distance in km, and lat_i and $long_i$ are the latitude and longitude at position i.

I then calculated excess allele sharing in distance bins using the method of Walter et al. (2015), by dividing the proportion of shared over total alleles at particular frequencies (from 2–7) by the proportion of individual pairs over all pairs in each distance bin. Counts of shared alleles between pairs were calculated using BCFtools by iterating over all possible pairs. To make this computationally feasible I restricted the counts to SNVs only. The significance of the distribution was assessed by regressing excess allele sharing across frequency categories on distance bin, treated as a continuous variable.

2.7.10 Meta-analysis

I combined the summary statistics of individual studies by meta-analysis, producing a weighted average from an augmented sample pool which was therefore better powered to detect associations. As these summary statistics were from studies of similar populations, many PCA-corrected, I used a fixed-effects, inverse-variance method to compute this average. This weights the i^{th} study with the inverse of its variance, δ_i^2, such that

$$W_i = \frac{1}{\delta_i^2} \tag{2.12}$$

The estimated overall effect β from n studies is given by

$$\beta = \frac{\sum_{i=1}^{n} \beta_i W_i}{\sum_{i=1}^{n} W_i} \tag{2.13}$$

where β_i is the effect size of the i^{th} study. The estimated overall variance δ^2 is given by

$$\delta^2 = \frac{1}{\sum_{i=1}^{n} W_i} \tag{2.14}$$

2.7.11 Measures of heterogeneity

$$Q = \sum_{i=1}^{n} W_i(\beta - \beta_i)^2 \tag{2.15}$$

Equation 2.15 defines Cochran's Q. W_i is the weight of the i^{th} of n studies (see equation 2.12), β_i is that study's effect size, and β is the overall effect size. Q follows a χ^2 distribution with $n-1$ degrees of freedom, allowing the generation of a P value of heterogenity. The Q statistic is made more intuitive by deriving from it I^2, the proportion of meta-analysis variance that is due to "truly" heterogeneous individual effect sizes (that is, not due to chance).

$$I^2 = \frac{Q - (n-1)}{Q} \tag{2.16}$$

Funnel plots allow a visual inspection of inter-study heterogeneity (Egger et al., 1997). They plot study effect size estimate against some measure of study precision, normally standard error. Since I was interested in the difference between the effect sizes reported by a set of smaller discovery studies and those of much larger validation sets, I used $\frac{1}{se}$, or precision, as my y-axis, as recommended by Sterne and Egger (2001). Normally, an asymmetric spread of small studies (which, due to their smaller precision, lie at the base of the curved funnel) is used to discern the underreporting of negative results, meaning publication bias. However, funnel plot skewing more accurately reflects heterogeneous estimates, and is agnostic to their cause.

The relationship between study size and estimate size is formally interrogated by Egger's test, a weighted regression of of effect size on se. Evidence for asymmetry is taken to be significant at $P < 0.1$ due to low power (Egger et al., 1997). This is subtly different from the significance of Cochran's Q, in that it tests for size-influenced rather than generalised heterogeneity.

2.7.12 eQTL

A quantitative trait locus (QTL) is a sequence in the genome whose variation partially contributes to the variation in some phenotype, or "quantitative trait". For "eQTLs",

this phenotype is the amount of mRNA (or protein) produced from a gene. GWAS SNPs are more likely to be eQTLs than one would expect by chance (Nicolae et al., 2010). It has become common to consider the overlap between GWAS loci and eQTLs when attempting to zero in on function.

I searched for eQTLs within TCGA data with a Ruby script that corrected for copy number variation (CNV) at each locus, as in Li et al. (2013). At each SNP position j I regressed gene expression (E) on average copy number within or spanning that gene's coordinates (CNV) at every gene (i) plus or minus 1Mb of the SNP:

$$E_i = CNV_i + \epsilon_i$$

I then regressed the unexplained variance in gene expression, ϵ, on the genotype at each SNP (G) under an additive model, leaving residual error term ω:

$$\epsilon_i = G_j + \omega_i$$

2.7.13 Survival tests

The Kaplan-Meier test is a nonparametric survival estimator which incorporates right-censored data, estimating survival as a function of time (Kaplan and Meier, 1958). Two Kaplan-Meier curves can be compared using the Mantel-Cox, or log-rank test. I used the Kaplan-Meier survival function estimator in chapter 4. Such survival curves cannot realistically be stratified on further covariates, however. For this, I additionally used the Cox proportional hazards method, which assumes that covariates are multiplicatively related to the hazard function. The hazard function (the probability of an event not having occurred at a particular time) is related to the survival function, and does not need to be explicitly defined in the Cox method (Cox, 1972). All survival analyses were performed in R using the `survival` package. Cox proportional hazards models in chapter 4 included age, sex, and cancer stage as covariates.

2.8 DNA

2.8.1 DNA extraction

2.8.1.1 From cultured cells

DNA was isolated from cell cultures with the DNeasy minikit (Qiagen) according to the manufacturer's instructions. The concentration and quality of the extracted DNA was assessed by quantification on the Nanodrop spectrophotometer (see §2.8.3).

2.8.1.2 From blood

Patient blood for GWAS cohorts was collected in EDTA tubes and stored at $-80°$C. After thawing the samples at room temperature, I extracted DNA using Maxwell®16 Blood DNA Purification kits (Promega UK Ltd) according to the manufacturer's instructions. Briefly, 420μl of blood and 230μl of elution buffer were added to each Maxwell®16 cartridge. The Maxwell®16 AS2000 instrument could then process 16 cartridges at a time, extracting DNA in a 30 minute run. Extracted DNA was quantified (see §2.8.3).

2.8.1.3 From bacteria

Bacteria grown for molecular cloning (see §2.8.9) had their DNA extracted in two ways depending on the amount of DNA required.

For smaller DNA harvests, single colonies were grown in small volumes of broth. Their DNA was extracted using the Nucleospin Plasmid Mini prep kit (Machery-Nagel), following the manufacturer's instructions. The process involves alkaline cell lysis, silica membrane adsorption in high-salt buffer, wash steps, and elution.

Mini prep kits work because DNA differentially adsorbs to silica depending on local hydrogen bonding. When chaotropic salt levels are high, water is unlikely to disrupt the interaction between DNA and a silica membrane; DNA will remain bound to the membrane while impurities are washed away. After a number of these washes, the DNA can be eluted in a low-salt buffer, which increases the hydrophobic effect at the DNA-silica barrier and removes the DNA.

To amplify positive clones, larger 200ml cultures were grown and processed with Qiagen's EndoFree Plasmid Maxi kit, according to their instructions. This kit resembles the Mini prep, but it selectively binds DNA in an anion exchange column, rather than to silica.

In this case, plasmid DNA binds the anion exchange resin under *low* salt and pH conditions, after treatment with an endotoxin removal buffer. Impurities are removed by medium salt washes. Finally, the DNA is eluted in a high-salt buffer and then concentrated and desalted by isopropanol precipitation.

2.8.2 Phenol:chloroform purification of DNA

Proteins and lipids partition preferentially into an equal volume phenol:chloroform mixture rather than into water; nucleic acids and salts do not. This fact allows DNA to be extracted from cell lysate, or purified of protein or lipid contaminants.

Eppendorf phase-lock gel tubes were centrifuged at 13,000rpm for 1min. 300μl of DNA product and 300μl of phenol:chloroform (Sigma) were added to each phase-lock tube, vortexed, and spun at 13,000rpm for 5mins. A further 300μl phenol:chloroform was added, spun as before, joined by 300μl of chloroform only, and centrifuged again. The upper aqueous phase was then transferred to a fresh eppendorf, mixed with 1μl glycogen (Roche), 270mM sodium acetate, and 66% ethanol, and incubated overnight at $-20°$C. This mixture was then spun at 15,000g for 30mins, washed with cold 70% ethanol, and spun again at 15,000g for 5mins. The precipitated DNA was relieved of its supernatant and resuspended in 50–100μl elution buffer (Qiagen) before quantification (§2.8.3).

2.8.3 Nanodrop DNA quantification

The Nanodrop 2000 microvolume UV-visible spectrophotometer (ThermoScientific) was routinely used to measure nucleic acid concentration. DNA and RNA maximally absord UV light with wavelength 260nm. After calibration with an appropriate blank control, 1μl of sample was loaded onto the instrument's optical pedestal and the 260nm-wavelength light absorbed by the sample was measured. The ratio of readings at 260nm

and 280nm gave an estimate of nucleic acid purity; a $\frac{260}{280}$ ratio of ~ 1.8 for DNA or ~ 2.0 for RNA was taken to evince acceptable purity. The appropriate constant of 50 and 40 for DNA and RNA respectively was used to obtain sample concentration.

2.8.4 PCR

Sequencing and RT-PCR primers were designed using Primer3 software (Untergasser et al., 2012). Stock primers were diluted to 100μM, with working stocks diluted at 1 : 10.

A typical 50μl PCR reaction contained 30–50ng of template DNA, 0.4pmoles of each primer, 2mM $MgCl_2$ (Promega), 1× Mg^{2+}-free PCR buffer (Promega), 0.2mM of each dNTP (Amersham), and 1 unit of Taq DNA polymerase (Bioline).

A normal thermocycler program (on PTC-255 machines, MJ Research) ran for 5mins at 95°C to denature DNA, and then underwent 35 cycles of 1min 94°, 1min 55°C, and 1min 72°C, ending with a final 10min extension step at 72°C. [Mg^{2+}], additives, and annealing temperature were varied to optimise amplifications with different primer pairs as necessary.

2.8.4.1 KASP genotyping

KASP genotyping (*k*ompetitive *a*llele-*s*pecific *P*CR, LGC) uses competing allele-specific fluorescence-activating primers to genotype SNPs. The competing forward primers ($\sim$ 20bp long, designed on PrimerPicker, KBioscience) differ at the position of the SNP, and incorporate a tail sequence complementary to a distinct Förster resonance energy transfer (FRET) casette: one primer binds a VIC-linked casette, the other a FAM-bound sequence. Both FRET casettes are present in KASP master mix; a common reverse primer must also be supplied. The binding and unquenching of a FAM or VIC oligo creates wavelength-specific fluorescence corresponding to the abundance of allele-specific amplicon. Thus, post-PCR relative FAM/VIC fluorescence, read on an ABI 7900HT Thermocyler qPCR machine, effectively genotypes the SNP.

KASP primers were diluted to 100μM and used to make an assay-ready mix, with volumes as listed in table 2.6. Per well, a KASP PCR required the master mix detailed

Primer	Volume (μl)
Allele 1	12
Allele 2	12
Common primer	30
H_2O	46

Table 2.6: Assay mix for KASP genotyping

Ingredient	Volume (μl)
Assay mix	0.14
KASP 2× master mix	5
H_2O	3.86

Table 2.7: Master mix for KASP PCR

in table 2.7. 9μl of master mix per well mix with the 1μl of input DNA or negative control. An example genotyping plot, generated by post-PCR reading on 7900HT, is shown in figure 2.1. The rs17816465 genotyping primers are listed in table 2.8.

2.8.5 SYBR

SYBR-Green Dye I binds to double-stranded DNA (dsDNA), quantifying amplification by producing fluorescence proportional to the amount of PCR product at each amplification step of qPCR. To perform SYBR-qPCR, I mixed 10ng DNA with 10μl FAST SYBR-Green mastermix (ThermoScientific), 300nM of forward and reverse primers, and H_2O to a reaction volume of 20μl. qPCR was carried out on the ABI 7900HT Thermocycler, with 20s of 95°C denaturation followed by 40 cycles of 1s at 95°C and 20s at 60°C.

To ensure that the dsDNA producing SYBR fluorescence was not from primer dimerisation or nonspecific amplification, I checked each pair of SYBR primers for non-specific PCR product formation by dissociation curve. After qPCR as described above, samples

name	5′–3′ sequence
rs17816465 A1	GAAGGTGACCAAGTTCATGCTAGGTACTTTTTATTGCAGGCTTCTGTT
rs17816465 A2	GAAGGTCGGAGTCAACGGATTGGTACTTTTTATTGCAGGCTTCTGTC
rs17816465 common	GCCACACAAAGGGAGGTTTCAAGTT

Table 2.8: KASP genotyping primers for rs17816465 genotyping. A1 and A2 are allele-specific primers which compete for the common reverse primer.

were denatured and slowly cooled to 50°C. An example satisfactorily sharp dissociation peak is shown in figure 2.2.

2.8.6 Sanger Sequencing

Individual PCR amplicons were sequenced by the Sanger method using the Big Dye Terminator v3.1 kit (Applied Biosystems). Briefly, 5μl of PCR product was incubated with 2μl of EXOSAP-IT (Affymetrix) for 15mins at 37°C followed by 15mins at 80°C. EXOSAP contains exonuclease I and alkaline phosphatase, which together digest unused dNTPs and primers from PCR. After this cleaning step, the sequencing reaction proceeded on a thermocycler, using 10μl 2× BDT reagent, 0.4pmoles of forward or reverse primer, and 20–50ng of PCR product, made up to 20μl. This reaction consisted of an initial 5mins denaturation step (90°C), followed by 35 cycles of 30s at 96°C, 45s at 50°C, and 4mins at 60°C. Unincorporated dye terminators were removed with the DyeEx 2.0 Spin kit (Qiagen) according to the manufacturer's instructions. The purified products were run on the ABI3730 DNA analyzer (Applied Biosystems) by the Oxford University Zoology department sequencing service, or sent to the Source BioScience Sanger sequencing service. Sequence traces were visualised using either A Plasmid Editor (ApE, by M. Wayne Davis), or SnapGene Viewer v3.3.3 (GSL Biotech). Primers used for sequencing in this thesis are shown in table 2.9.

2.8.7 MSI status

The microsatellite stability of the homozygous *MSH3* mutant in chapter 6 was assessed by the Oxford Medical Genetics Laboratory at the Churchill Hospital, using Promega's MSI Analysis System, version 1.2, according to manufacturer's instructions. The sample would have been considered microsatellite unstable (MSI) if two or more of the five mononucleotide markers (BAT-25, BAT-26, NR-24, NR-21, and MONO-27) had been unstable.

Elevated microsatellite instability at selected tetranucleotide repeats (EMAST) status was derived using the five primer pairs in table 2.10, where each forward primer was HEX-tagged and sized by comparison to a 500bp LIZ-tagged internal lane standard.

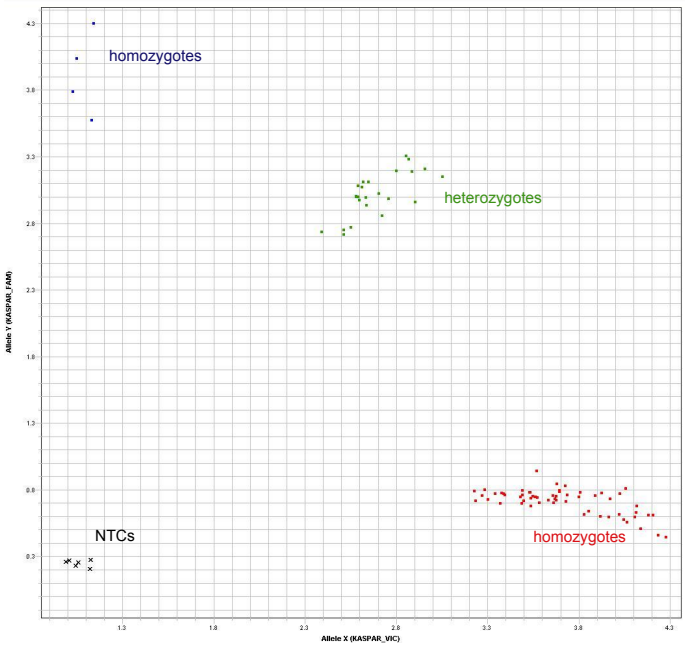


Figure 2.1: A genotyping cluster plot after KASP genotyping PCR. This example was performed on a CORGIbcd plate with primers designed against rs17816465. Green, heterozygous samples; blue, minor allele homozygotes; red, major allele homozygotes; black, negative controls (water). The axes measure post-read fluorescence of VIC and FAM.

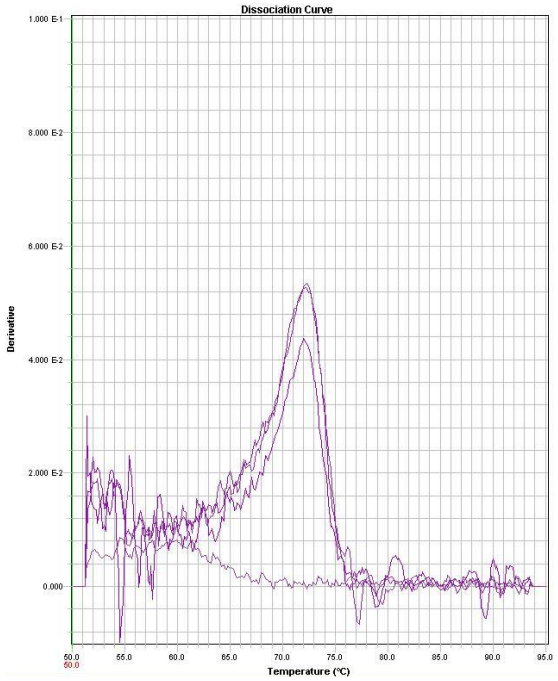


Figure 2.2: SYBR-Green primer dissociation curve showing a single, discrete peak in three technical replicates, with no fluorescence in the negative control.

chr	gene	target	forward	reverse
1	*MUTYH*	p.176	AAGACACCCAAGACTCCTGG	GGTGGGTGTAGAGAAGGCTT
1	*MUTYH*	p.393	TCACTTACCTCCCCAAGGTG	AAATTCTGCTGGTGCAGAGG
2	*MSH2*	47641560:A>T	CATCACTGTCTGCGGTAATCA	GCCATTTAAAGCTAGTTATCTAATCCA
2	*MSH2*	p.680	CCTACGCGATTAATCATCAGTG	ACATTTCAGCCATGAACGTG
2	*MSH6*	p.686	GCACGAGTGGAACAGACTGA	AGCACCATTCGTTGATAGGC
2	*MSH6*	p.1035	GGGATCATGGAAGAAGTTGC	TACAGCTGGCAAACAGCACT
2	*MSH6*	p.1307	AGGCATGCATGGTAGAAAATG	CATCATCCCTTCCCCTTTTA
3	*MLH1*	p.22	CACTGAGGTGATTGGCTGAA	CGGCCCGTTAAGTCGTAG
3	*MLH1*	p.561	CCACAGCCAGGCAGAACTAT	TTTTCAGAAACGATCAGTTGAA
3	*MLH1*	p.618	GCTGGACAGAGGAAGATGGT	ATGGCTGTCACACCTCATCA
3	*MLH1*	37090394:G>C	CAGAGTGGCAGATAGGAGCA	CTAGTCCTGGGGTGCCAGT
3	*MLH1*	p.751	TTCCAAACTCCTGGAAGTGG	GGCAAGTATAAGTCTTAAGTGCTACC
5	*APC*	p.83	GGGAAAGTCCTAAATAATTTTTGTT	CCCTCTTCTTGGAAATGAACC
5	*APC*	112163705:T>C	GCTTGGCTTCAAGTTGTCTTTT	GAAGTGGGAGGATTGCTTGA
5	*APC*	p.653	AAAGGAGATGTGGAATACTTGGA	AGCTGCAGCACTTCCCATAG
5	*APC*	p.2485	ACGCCAGTCAACTTTCATCA	TTCAGAATGAGACCGTGCAA
5	*MSH3*	p.268	TCACTGGAACTGCCCAAAAT	TGGTAGGTAGGCTTCAATTGG
7	*PMS2*	p.1	ATTTCCAGGGAGGTTGGAAT	GAGAGACGGGAAGTATTTTTGC
7	*PMS2*	p.46	CAACAACATTCACAGATCATTTCTT	CAGAACCTGCTAAGGCCATC
7	*PMS2*	p.205	GTTGAAGTAACCGGCCATCA	TATGCCAAAATGGTCCAGGT
7	*PMS2*	p.649	CAGAGCAAGACTCTGTCTCAAAA	GACATGTCAGCCTCTCAGGTT
7	*PMS2*	6022454:C>T	CTCAAACTCCTGGCCTCTTG	TTGGTCAGTTTAACCTGGGATT
10	*BMPR1A*	p.48	TTGTCACGAAACAATGAGCTTT	AACTCTTAAGAAGGGCTGCAT
10	*BMPR1A*	p.155	AGCTACGCCGGACAATAGAA	CTGGCTTTACAAACAGCGGT
10	*BMPR1A*	p.309	TGAGCATTACTTCTCCCTAGCC	CTTTCCTTGGGTGCCATAAA
10	*BMPR1A*	p.443	TGATGTGCCCTTGAATACCA	TGTGCCCACAAAAATTAAAAA
12	*POLE*	p.366	CTCAGACCCAGGGAGGAAC	GCAGACCTCTGACTGCTGTG
12	*POLE*	p.424	AGCCGGGATGTGGCTTAC	TTGCATCTGTCTGTGTGGTG
15	*FMN1*	rs17816465	TGGACTTCTATGAAAGGGGAAAA	AGATTTCCCCTTTGCCCTCT
15	*GREM1*	rs1406389	GTTCTTTCAGCCCAGATGCC	ACCTACCGACGTGTAAAACCA
17	*RNF43*	p.157	TCAATCCTCACATGGGCCTT	ATTTGCGTGTTCAAGGTGGG
18	*SMAD4*	48603147:G>A	GATTTGCGTCAGTGTCATCG	TCAAAAATGTCATCATCCCAGT
19	*POLD1*	p.478	GTGTGTCCCTGTCCTTGGAA	GGCGTGAAGTTGAGGTCAGA

Table 2.9: Sanger sequencing primers used in this thesis. Primer sequences are given 5′–3′; target if prefaced by "p." refers to protein position.

marker	location	5′–3′ primers	product size
D20S82	20p12.3	GCCTTGATCACACCACTACA GTGGTCACTAAAGTTTCTGCT	246–270bp
UT5037	chr8	TTCCTGTGAACCATTAGGTCA GGGAGACAGAGCAAGACTC	145bp
D2S443	2p13.2–2p13.1	GAGAGGGCAAGACTTGGAAG ATGGAAGAGCGTTCTAAAACA	251bp
D21S1436	21q21.1	AGGAAAGAGAAAGAAAGGAAGG TATATGATGAAAGTATATTGGGGG	178bp
D9S747	9q32	GCCATTATTGACTCTGGAAAAGAC CAGGCTCTCAAAATATGAACAAAAT	182–202bp

Table 2.10: Primers for amplifying EMAST markers.

The sizing of PCR amplicons was performed on an ABI 3730 Capillary Electrophoresis machine by the Oxford Medical Genetics Laboratory.

2.8.8 Chromatin conformation capture

Chromatin conformation capture (3C) detects the looping together of distant regions of DNA. Cells are cross-linked to preserve DNA topology, and their genome is sheared by restriction enzymes. The fragments are ligated and cross-linking is reversed. In this way, small ligation products are formed by those restriction fragments that were physically close together in the cells: amplification across the ligation boundaries shows looping.

Cells ($\sim 10^7$.sample^{-1}) were cross-linked in 1% formaldehyde in phosphate buffered saline solution (PBS, Sigma-Aldrich) *in situ* for 10mins at room temperature. They were transferred to ice, and cold 0.125M glycine-PBS was used to wash the cells and to quench the residual formaldehyde. Cells were scraped into fresh glycine-PBS and spun down at 400g. After a final PBS wash and spin, the cells were lysed in ice-cold buffer (10mM Tris pH 7.5, 10nM NaCl, 0.2% IGEPAL, protease inhibitors (Complete mini, Roche)) for 30mins. Lysis was completed in a Dounce column. 800g centrifugation pelleted the nuclei, which were resuspended in 1.2× restriction enzyme buffer (Buffer 2.1, NEB). 0.3% SDS was added to lyse the nuclei and solubilise the chromatin within. After 30mins at 37°C this was sequestered by the addition of 2% Triton-X for another 30mins at 37°C, before 1500U of restriction enzyme (in my case EcoRI-HF; see table 2.11) were added. After 4 hours another 1500U of restriction enzyme were added to digest the DNA overnight.

The restriction enzyme was inactivated by the addition of SDS (1.6%) at 70°C for 20mins. The digest was then transferred to 50ml tubes and diluted into 8ml water with 1% Triton-X, 1× ligation buffer (50mM Tris pH 7.5, 10mM DTT, 10mM $MgCl_2$), 100ng.ml^{-1} BSA, and 1mM ATP. After an hour's incubation at 37°C (to quench the SDS), 4000U ligase were added at 16°C. 4 hours later another 4000U ligase were added, and the ligation was left to incubate for approximately 18 hours.

The ligation reaction was warmed to room temperature over 30mins, and then 300μg of Proteinase K (Qiagen) were added. Cross-links were reversed overnight in a 65°C waterbath. To reduce contamination, a 45mins 37°C incubation with 300μg RNAse A (ThermoScientific) followed.

The un-crosslinked DNA was then cleaned by phenol-chloroform extraction (§2.8.2) and ethanol precipitation. The DNA pellet was resuspended in 10mM Tris pH 7.5 and re-purified in Amicon Ultra Centrifugal Filters (0.5ml; Millipore), flushing contaminants through with three 1×TE washes. The 3C library produced was then quantified by both Nanodrop (see §2.8.3) and by SYBR-qPCR (§2.8.5) against known concentrations of genomic DNA with *GAPDH* probes internal to restriction digestion sites. Once 3C concentration was known, the libraries could be analysed by qPCR, normalised to crosslinking frequencies for the cloned bacterial artificial chromosome (BAC) RP11-10I10 (BACPAC Resources), with optimised primers spanning a custom 6FAM-tagged Taqman probe (5′-ACGACAATCTTTAAGTTTC-3′).

2.8.9 Cloning

The BAC used for 3C was grown on LB agar with 12.5μg/ml chloramphenicol at 37°C, the DNA extracted and expanded by Qiagen's Maxi kit (§2.8.1.3).

All other plasmids were cloned in XL-10 Gold UltraCompetent cells (Stratagene). The bacteria were transformed in 14ml BD Falcon polypropylene round-bottomed tubes by adding 10μl DNA, incubating on ice for 30min, and then heat-shocking the cells at 42°C for 30s. After an hour of recovery in LB medium at 37°C with agitation, transformed bacteria were plated LB agar with 100μg/ml ampicillin.

The sequence of the cloned BAC was checked with multiple restriction digests, using

Enzyme	Catalogue number
AgeI	R0552L
BglII	R0144L
BbsI	R0539L
EcoRI-HF	R3101M
XmnI	R0194L

Table 2.11: Restriction enzymes used. All enzymes bought from New England Biotech.

the enzymes in table 2.11.

2.8.9.1 Gel electrophoresis

PCR, ChIP sonications, and cloning products were analysed using agarose gel electrophoresis. The 1–2% agarose gels were stained with ethidium bromide and DNA was run alongside a 1kb DNA ladder using Tris-Borate EDTA running buffer. Gels were visualised on a Universal III Gel Doc (Bio-Rad).

2.9 RNA

2.9.1 RNA extraction from cultured cells

RNA was isolated from cell cultures with the RNeasy minikit (Qiagen) according to the manufacturer's instructions. The concentration and quality of the extracted RNA was assessed by quantification on the Nanodrop spectrophotometer (see §2.8.3) and by gel electrophoresis (see §2.8.9.1). Two strong bands at 5kb and 2kb corresponded to 28S and 18S respectively, the two most abundant rRNA species. The strength and precision of these bands was taken as a subjective indication of the quality of the less concentrated mRNA population.

2.9.2 cDNA creation

Before reverse transcription, extracted RNA was treated with DNase I (ThermoScientific) to minimise RT-PCR-confounding contamination. The DNase I was subsequently inactivated by EDTA (ThermoScientific).

Complementary DNA (cDNA) was reverse transcribed *in vitro* from purified RNA

using the High Capacity cDNA Reverse Transcription Kit (Applied Biosystems) according to the manufacturer's instructions. Random hexamers primed the Reverse Transcriptase.

2.9.3 qRT-PCR

The ABI 7900HT cycler (Applied Biosystems) threshold was manually set at 0.2 and standard protocols were used to calculate relative expression with the $\Delta\Delta C_T$ method described by Livak and Schmittgen (2001), using *GAPDH* as an endogenous control. Pre-prepared Taqman gene expression assays were used, which contain primer pairs and probes optimised by Applied Biosystems. The Taqman qRT-PCR primers used in this thesis were against *GREM1* (Hs00171951), *ID1* (Hs03676575), and *GAPDH* (Hs99999905). A typical reaction contained, per 10μl well, 10ng of purified cDNA, 0.5μl of 20× Taqman gene expression assay, 5μl of 2× Taqman FAST Universal Mastermix, and H_2O. Samples were run in triplicate on the 7900HT cycler for an initial 20s of 95°C denaturation, followed by 20 cycles of 1s 95°C and 20s of 60°C.

2.10 Protein

2.10.1 Western blotting

1CT cell pellets were lysed for 20 minutes on ice in RIPA buffer (Tris-Hcl 50 mM, NaCl 150mM, 0.5% sodium deoxycholate, 1% Triton-X, 0.1% SDS, 5mM EDTA, pH 7.4) with the addition of protease inhibitors (Complete mini, Roche) and phosphatase inhibitors if necessary (PhosSTOP, Roche). The samples were then centrifuged at full speed for 2 minutes on a tabletop centrifuge to precipitate cell debris.

Lysates' protein concentrations were calculated using the Pierce bicinchoninic acid (BCA) Protein Assay Kit (ThermoScientific). The stock BCA solution contains copper(II) sulfate, the Cu^{2+} ions of which are reduced to Cu^{+} in an amount proportional to the concentration of protein present. Cu^{+} complexes with BCA to form a purple product that absorbs light at 562nm. Hence, the spectrophotometric absorbance at 562nm of samples after the assay gave a measure of their protein concentration. To

quantify this more rigourously, I used a standard curve of diluted bovine serum albumin. BCA reagents A and B were mixed in a 1 : 50 ratio, and 1ml was added to diluted cell lysates for 30mins at 37°C. 562nm-wavelength light absorbance was read on a spectrophotometer (NovaspecII, Pharmacia Biotech). The known concentrations of the albumin standards were used to calculate the unknown lysate concentrations, which were then diluted in buffer to provide a constant protein load per well, normally 35μg.

Each lysate was mixed with 4× loading dye (Invitrogen) and denatured at 95°C for 10 minutes. Blotting followed the manufacturer's instructions for the NuPAGE Gel system (Invitrogen). The denatured lysates were run on a 12% NuPAGE Novex Bis-Tris Mini Gel at 120V for roughly 2 hours, alongside a protein ladder (Spectra Multi-colour, Fermentas). 10μl of antioxidant were added to the running buffer in the small chamber of the gel tank. Subsequent blotting onto a PVDF membrane (Immobilon P, Millipore) in a semi-dry tank carrying 120mA took approximately 2.5 hours.

The membranes spent 1 hour rocking at room temperature in 10% TBS-milk (Marvel) or TBS-BSA to block. They were then incubated overnight in primary antibody-containing PBS or 5% TBS-BSA at 4°C. The primary antibodies used in this thesis were rabbit anti-SMAD1/5/9 (Cell Signaling #9511; 1 : 1000) and mouse anti-SMAD1/5 (Bio Matrix Research, BMR00479; 1 : 1000). After washing in TBS-tween, the membranes were incubated for 1 hour at room temperature with the appropriate secondary antibody, and often with an HRP-conjugated anti-actin antibody (as loading control; SantaCruz sc-47778 HRP; 1 : 10, 000), again in PBS, TBS-milk or TBS-BSA. Secondary antibodies used were goat anti-mouse (Dako P0447, 1 : 2000) and goat anti-rabbit (Dako P0448, 1 : 2000).

After further washes, the membranes were submerged in ECL reagents (GE Healthcare), and chemiluminescence was detected either with chemiluminescent film (GE Healthcare) using Xograph's Compact X4 processor, or with ChemiDoc MP Imaging in a Universal Hood III (Bio-Rad).

2.10.2 Chromatin immunoprecipitation

Chromatin immunoprecipitation (ChIP) provides evidence of interaction between antibody-bound protein and DNA fragments. I used cross-linked cells, and broke up the genomic DNA using sonication. Throughout the immunoprecipitation I used protease inhibitors (phenylmethylsulfonyl fluoride, PMSF, Sigma-Aldrich, and Complete mini, Roche) and the histone deacetylase inhibitor sodium butyrate (Sigma-Aldrich) to protect against digestion.

ChIP was performed on CRC cell lines at 70% confluence in 15cm petri dishes. The cell lines HCD57, LS180, COLO320, SW1222, and T84 were used. 1% formaldehyde (Sigma-Aldrich) was added to each dish of cells at room temperature, and rocked for 10mins to cross-link protein to DNA. 115mM glycine (Sigma-Aldrich) was then added to quench the formaldehyde for 5mins. The cells were twice rinsed with 4°C PBS supplemented with 0.1% PMSF and 0.2% sodium butyrate. The cells were scraped from their dishes into 15ml tubes and spun at 700g for 5mins, before resuspension in ice-cold lysis buffer (1% sodium dodecyl sulfate (SDS, Sigma-Aldrich), 10mM EDTA (Invitrogen), and 50mM pH 8.0 Tris-HCl (Invitrogen)). After 10mins incubation, the lysed cells were sonicated using the Bioruptor Plus (Diagenode) in 7–15 15s cycles to generate DNA fragments of 1000–1500bp. After sonication, the cells were centrifuged at 700g for 10mins, with a small amount of supernatant run on a 1% agarose gel to confirm fragment size. The remaining supernatant underwent 1 : 10 dilution in dilution buffer (1% Triton-X-100 (Sigma-Aldrich), 2mM EDTA, 20mM pH 8.0 Tris-HCl, and 150mM sodium chloride, with Complete mini protease inhibitor, 0.1% PMSF, and 0.2% sodium butyrate) and was then separated into aliquots: one "input" chromatin, one "no antibody" control, and one tube for each immunoprecipitation, to which 5μl antibody were added. The antibodies I used in chapter 3 were anti-dimethyl-histone H3K4 (Millipore 07-030), and anti-acetyl-histone H4 (Millipore 06-866). The antibody used for ChIP-seq in chapter 6 was mouse anti-SMAD1/5 (Bio Matrix Research BMR00479).

The antibody and control tubes were rocked overnight at 4°C to allow adequate antibody-protein binding. The input tube was merely kept refrigerated. The next day, 5μl protein A dynabeads (ThermoScientific, 10001D) were added to each rocked

tube and incubated for 4 hours. A series of 1ml ice-cold washes were performed using, sequentially:

1. Tris/Sucrose/EDTA (TSE) I (1% Triton-X-100, 2mM EDTA, 20mM pH 8.0 Tris-HCl, 150mM NaCl, 0.1% SDS). On/off wash.
2. TSE II (1% Triton-X-100, 2mM EDTA, 20mM pH 8.0 Tris-HCl, 500mM NaCl, 0.1% SDS). Four washes: on/off, 5mins, 2 × 10mins.
3. Buffer III (0.25M lithium chloride, 1mM EDTA, 10mM pH 8.0 Tris-HCl, 1% Igepal (Sigma-Aldrich), 1% sodium deoxycholate (Sigma-Aldrich). On/off wash.
4. 1× Tris-EDTA. Three washes: on/off, 5mins, 10mins.

Each wash buffer contained PMSF and sodium butyrate as above. A MagnaRack magnetic separation rack (Invitrogen CS15000) was used to retain the dynabeads between washes.

30mins of room temperature incubation in 300μl extraction solution (1% SDS, 0.1M sodium bicarbonate) unbound the antibody from the dynabeads. Then NaCl was added to 0.7M and reverse cross-linking occurred overnight at 65°C. QIAquick PCR purification columns (Qiagen) removed the now unbound protein, leaving PCR-ready DNA.

ChIP product was quantified by SYBR-qPCR (§2.8.5) in biological and technical triplicate, using input chromatin and the *RHO* promoter as negative controls.

2.11 Cell culture

2.11.1 Maintenance of cell lines

All cell culture work was performed in a class II laminar flow hood. Cell lines were grown from frozen stocks, thawed at 37°C, pelleted, and resuspended in the appropriate tissue culture medium. They then grew in incubators at 37°C with 5% CO_2. Cells were passaged before reaching confluence; trypsin-EDTA (Invitrogen) was used to detach them to avoid selective subculture. For cancer cell lines, the culture medium was either Dulbecco's Modified Eagle Medium (DMEM) with L-glutamine (Lonza), or Roswell Park Memorial Institute 1640 (RPMI-1640) medium (Sigma-Aldrich), in

each case supplemented with 10% fœtal bovine serum (50μg/ml FBS, Gibco) and 1% penicillin-streptomycin (50μg/ml, Gibco).

The exception was the immortalised epithelial cell line HCEC-1CT (1CT hereafter; Roig et al., 2010). 1CTs were grown in a low-O_2 incubator in custom medium. This consisted of, per 100ml:

- 100μl apotransferrin (0.002g.(ml H_2O)$^{-1}$)
- 20μl EGF (100μg.ml^{-1})
- 100μl hydrocortisone (1mg.ml^{-1})
- 500μl insulin (2mg.ml^{-1})
- 100μl sodium selenite (5μM)
- 100μl gentomycin sulfate (50mg.ml^{-1})
- 2000μl cosmic calf serum
- 77.7ml DMEM
- 19.4ml M-199

As soon as possible after thawing, cell subcultures undergoing log-phase growth were returned to liquid nitrogen storage. In preparation, cells were detached and then resuspended in a freezing medium consisting of 90% FBS, 10% DMSO.

1CT cells were treated with recombinant human BMP2 (355-BM-010, R&D Systems) or BMP4 (314-BP-010, R&D Systems).

2.11.2 Cell counting

Cell numbers were counted in two ways. The first was to use a FastRead102 disposable hæmocytometer (Immune Systems), using the calculation:

$$\text{cells.ml}^{-1} = \frac{\text{Cells counted}}{4 \times 4 \text{ grids counted}} \times 10^4$$

The second method was to use Invitrogen's Countess II automated cell counter, according to their instructions.

Chapter 3

A New Colorectal Cancer Risk-Associated Variant Near *GREM1*

3.1 Introduction

3.1.1 *GREM1*

In 1998, Hsu et al. found a synthetic mRNA whose ventral injection into *Xenopus* frog embryos caused the formation of a secondary neural axis. They named this novel signalling gene *gremlin* (*GREM1* in humans). The small protein it encoded was found to bind and antagonise BMP2 in the manner of known TGF-β inhibitors *noggin* (*NOG* in humans), *chordin* (*CHRD*), and *follistatin* (*FST*). However, *gremlin* displayed no sequence homology to these genes. Instead, it shared a cystine knot motif with the secreted proteins Cerberus and DAN (differential screening selected gene abberative in neuroblastoma). This novel family was called the DAN family after its founding member; it has since grown to include NBL1 (DAN), GREM1, GREM2 (also known as PRDC), SOST, USAG1, DAND5 (also known as COCO), and CER1 (reviewed in Nolan and Thompson, 2014).

Some DAN family antagonists have an unpaired cysteine residue near their cystine knot, which was thought to enable dimerisation by disulfide bonding, as in many TGF-β ligands (Avsian-Kretchmer and Hsueh, 2004). The molecules studied, DAN and GREM2, do indeed form stable dimers, but these are noncovalent. The unpaired cysteine is not necessary for dimerisation (Kattamuri et al., 2012). Noncovalent GREM2 dimers

bind their ligand in aggregating chains, with each component monomer able to bind a BMP dimer. This both sequesters extracellular BMPs, and blocks their interaction with type I BMP receptors (Nolan et al., 2016). It is likely that GREM1 oligomerises and acts in the same way.

GREM1 is important in many signalling processes. Not least, it is a crucial member of the regulatory circuitry of the developing limb bud. Early limb bud development relies on the positive feedback loop established between an *FGF4*-expressing apical ectodermal ridge (AER) and an *SHH*-expressing zone of polarising activity (ZPA) on the posterior mesenchyme. Woychik et al. (1985) generated by insertional mutagenesis a mouse with severe limb developmental defects, termed *limb deformity* (*ld*). The *ld* locus was mapped to a novel set of transcripts for the gene *Formin* (Woychik et al., 1990), and showed disruption of this vital dialogue between the AER and the ZPA. Zúñiga et al. (1999) clarified the *ld* pathology: not *Formin*, but *Gremlin* expression is lacking in the posteror limb bud mesenchyme of affected mice. A *Grem1*-expressing graft rescues the phenotype. SHH in the ZPA relays its signal through GREM1, to inhibit BMPs in the AER and to disinhibit FGF4, which in turn upregulates SHH. This allows normal limb development.

GREM1 is also active in the human gut. The protein helps to reinforce the signalling gradients of the colon's crypt-villus axis. At the bottom of colonic crypts, a high-WNT, low-BMP environment nurtures and encourages the tissue's replenishing stem cell population (Hardwick et al., 2004). This niche benefits from the upward diffusion of GREM1, which antagonises pro-apoptosis, pro-differentiation BMPs. As shown in figure 1.3, BMP signal negatively regulates intestinal stem cell renewal through SMAD intermediates (Qi et al., 2017). The healthy crypt epithelium does not express *GREM1*; instead, subepithelial myofibroblasts secrete the antagonist (Kosinski et al., 2007). This can be seen by *in situ* hybridisation in figure 3.1.

This expression pattern becomes deranged in hereditary mixed polyposis syndrome (HMPS). Instead of subepithelial myofibroblasts, the colon epithelium itself expresses *GREM1*, in vastly increased amounts (figure 3.2, Jaeger et al., 2012). A mouse model of this defect, overexpressing *Grem1* with an intestinal epithelium driver, develops ectopic

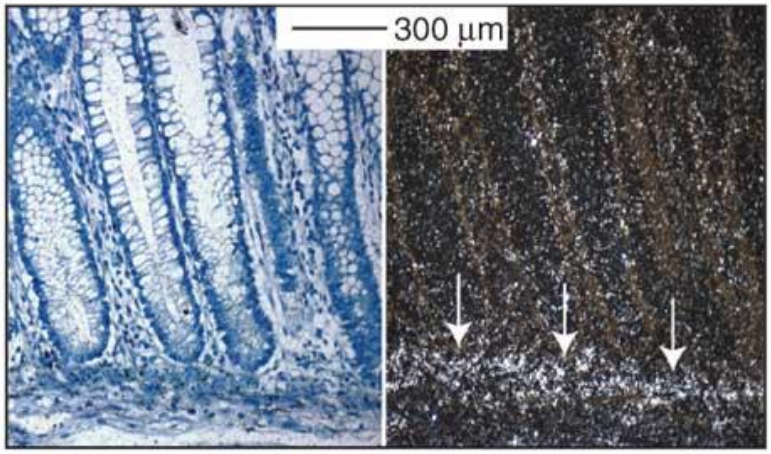


Figure 3.1: *GREM1* is expressed exclusively by colonic subepithelial myofibroblasts in healthy individuals. ISH from Jaeger et al. (2012)

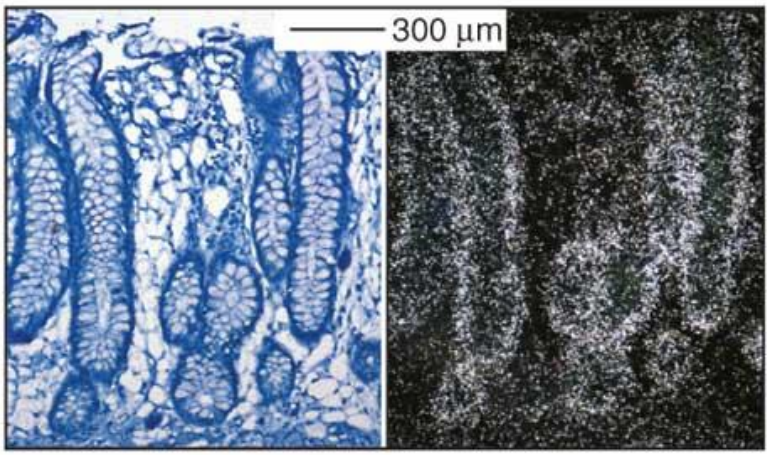


Figure 3.2: In HMPS, *GREM1* expression is ectopic and vastly increased. ISH from Jaeger et al. (2012)

crypts and intestinal polyps (Davis et al., 2015).

3.1.2 Previous *GREM1* GWAS signals

I described in §1.4.3 how the *GREM1* locus was associated with sporadic CRC risk during the search for the HMPS mutation. The original 2008 *GREM1* GWAS signal was subsequently refined. As with any GWAS follow-up, this meant denser genotyping and larger studies.

Three times, regional fine-mapping of association at *GREM1* was performed as part of a genome-wide sweep of CRC-associated loci. Whiffin et al. (2013) re-tested these loci after imputing genotypes in 5626 cases and 7817 controls, using as reference a combination of 1000 Genomes Project (1000G) phase 1 sequences and a smaller 253 "CG" panel, which was highly enriched for early-onset CRC patients. The SNPs with the smallest P values of association after meta-analysis were taken to be the best candidates for causal variation at each region. Near *GREM1*, this SNP was rs1406389 ($P = 7.57 \times 10^{-12}$), beating the original tag rs4779584, to which it is correlated ($r^2 = 0.67$, $D' = 0.82$). In 2014, Whiffin et al. essentially repeated this analysis with an updated 1000G reference panel. This came up with a different "best" SNP at 15q13: rs73376930 ($P = 1.12 \times 10^{-11}$). Finally, Du et al. (2016) compared 11,900 cases and 14,311 control allele frequencies imputed on a 1000G reference, and preferred rs2293582 ($P = 10^{-9}$).

The slightly crude reporting of smaller and smaller P values is perfectly acceptable, of course, and the two Whiffin et al. papers were in any case not particularly interested in the *GREM1* locus, having broader aims and hypotheses. But what these papers gained in scope they somewhat lost in detail, for they ignored existing work suggesting that the 15q13 GWAS signal derives from more than one variant.

Back in 2011, only 14 CRC-associated loci had been identified by GWAS. Three were close to BMP pathway member genes: rs4779584, rs4444235, and rs961253 lay near *GREM1*, *BMP4* and *BMP2*, respectively. Tomlinson et al. (2011) proposed to use BMP pathway member gene loci as candidate regions, which could be examined for more association using a lower statistical threshold. The authors genotyped 442 SNPs around the three genes in roughly 10,000 case-controls, filling in genotypes with imputation.

None of the new SNPs was more strongly associated than the original tags, except for one near *GREM1*. rs16969681 had a stronger signal than the incumbent rs4779584.

Had Tomlinson et al. (2011) been following the methodology of Whiffin et al. (2013, 2014), rs16969681's improved post-validation P value ($P = 5.33 \times 10^{-8}$) would have superceded the initial tag, with no further exploration. However, the authors instead made a point of using not only base P value but also other statistics to evaluate their risk model. In particular, in a rs16969681-rs4779584 haplotype test, the rs4779584 risk allele remained risk-associated regardless of the other SNP's genotype. Though the SNPs were correlated ($r^2 = 0.29$, $D' = 0.76$), and therefore non-independent as covariates in linear tests, it was clear that rs16969681, as well as "beating" the tag SNP to a smaller overall P value, nonetheless also failed to fully account for the original signal. Reverse stepwise logistic regression with other regional candidates produced a two-SNP model of *GREM1* CRC risk, formed of rs16969681 and rs11632715. Both SNPs were individually strongly associated with risk, remained independent as covariates, and produced a model with a lower Aikake's information criterion (AIC) than regression on rs4779584 alone.

The case for multiple causal signals at *GREM1* (themselves tagged by multiple SNPs) was thus subtle, and relied on exploratory, descriptive statistics. Perhaps it is understandable that later CRC GWAS fine-mapping largely ignored the argument. However, this chapter will counterargue three points. First, the case for rs16969681's functional effects has now been conclusively made, lending strong weight to the reasoning of Tomlinson et al. (2011). Second, my own pursuit of similar lines of thought has led to the discovery of a third genome-wide significant SNP in the region, providing a demonstration of the usefulness of this approach. And third, more speculatively, I will contend that detailed genetic deconvolution of a GWAS tag is a more reliable route to the discovery of functional variation than more high-throughput methods. As such I will both restate and develop the case of Tomlinson et al. (2011) in order to offer a more accurate description of CRC risk genetics at *GREM1*.

3.2 Results

3.2.1 Fine-mapping risk near *GREM1*

As discussed in §2.6.3, imputation fills in missing data with statistically inferred genotypes. The variant types to impute are provided by a dense reference panel, or combination of reference panels.

In the megabase region surrounding *GREM1*, I imputed the genotypes of seven case-control cohorts against a combined reference panel of 1000G (phase 1, integrated version 3, March 2012) and 196 30×-deep genome sequences of early-onset CRC cases (the "CGI196" panel, see §2.4.4). I used IMPUTE2 with parameters as specified in §2.6.3. The seven cohorts are described in detail in §2.2 and summarised here in table 3.1.

Cohort	Cases	Controls	Total
CFR1	1175	999	2174
CFR2	795	2242	3037
LP1	888	899	1787
LP2	2659	2798	5457
SP1/MD1	973	998	1971
SP2/MD2	2007	2075	4082
VQ58	1794	2686	4480
All	11,063	11,925	22,988

Table 3.1: Summary of imputed fine-mapping discovery cohorts.

After file conversion and filtering, the genotypes were tested for CRC association by SNPTEST, conditioning on sex. These naïve tests were then meta-analysed using META. The combined discovery cohorts had 75% power to detect associated SNPs with control MAF= 0.2, $OR = 1.15$ at a genome-wide significant level ($P < 5 \times 10^{-8}$; see figure 3.3).

This workflow, described in more detail in §2.6.3, deviates very little from most standard fine-mapping protocols. Of note however are the use of a combined reference panel enriched (presumably) for European CRC-causing variants, and a particularly stringent info score cutoff for imputation output (info ≥ 0.9). The results of the basic meta-analysis are shown in figure 3.4.

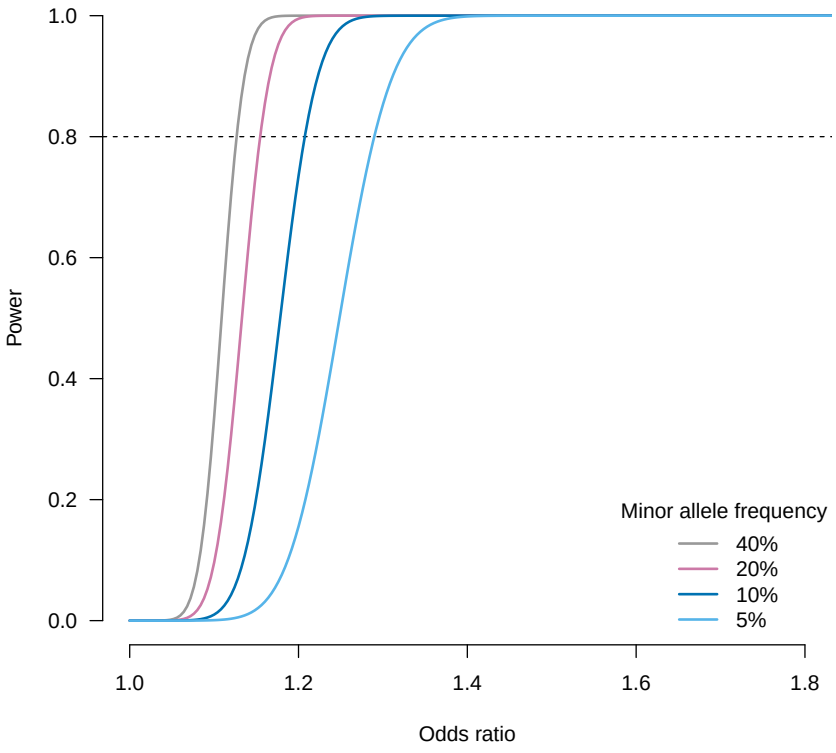


Figure 3.3: Estimated power to detect SNP associations within the combined discovery cohorts at $P < 5 \times 10^{-8}$ given different control minor allele frequencies.

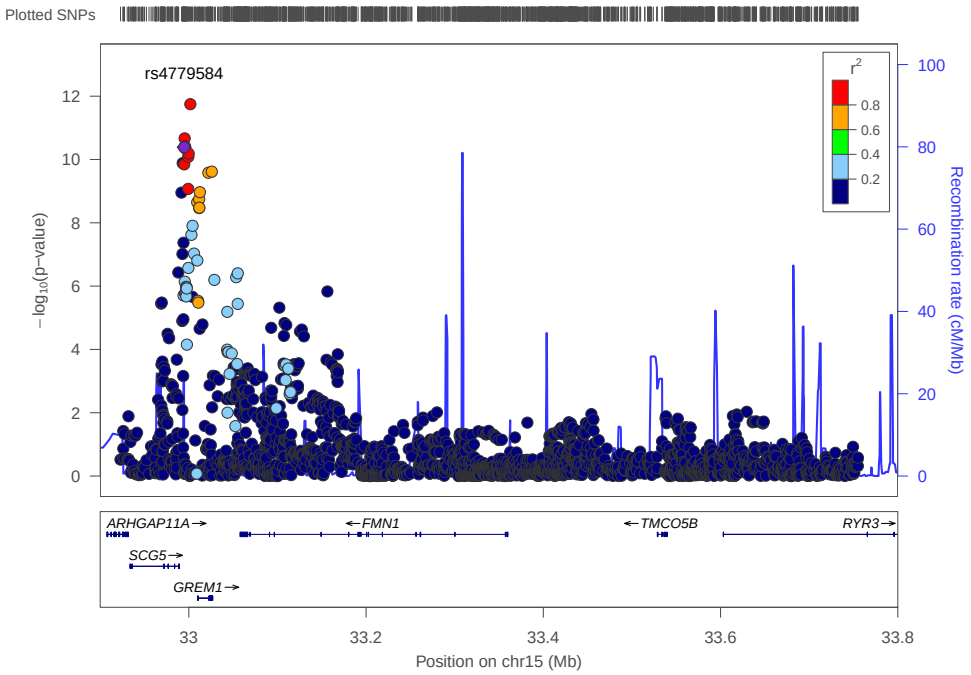


Figure 3.4: A Locuszoom (Pruim et al., 2010) plot summarising the first association testing and meta-analysis of the post-imputation discovery set. Each dot is a SNP; colour marks LD with rs4779584, which is purple.

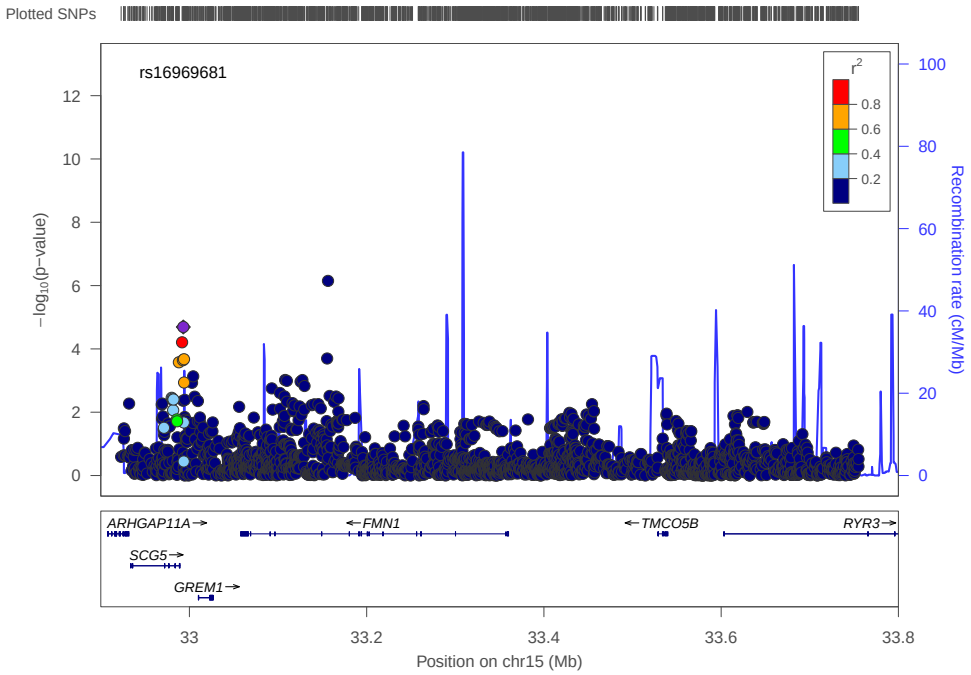


Figure 3.5: A Locuszoom plot showing the association landscape after conditioning on rs4779584. The second smallest P value is for rs16969681, which is marked in purple.

The top SNP from the first association test was rs58658771 ($P = 1.79 \times 10^{-12}$; $I^2 = 38.1\%$). This SNP is in strong LD with rs4779584 ($r^2 = 0.766$, $D' = 0.932$) and represents essentially the same signal. However, instead of reporting yet another marginal improvement as a possible candidate, I wanted to entertain the hypothesis that rs4779584/rs58658771 tagged multiple SNPs.

The conclusion of Tomlinson et al. (2011) was that rs4779584 incompletely captured other regional risk associations. The converse statement therefore suggested itself: that the underlying risk signals would remain incompletely associated with CRC after conditioning on rs4779584. To the degree that they were independent of rs4779584, the contributing risk SNPs' test statistics should remain large after conditional analysis.

I therefore used SNPTEST to test each cohort, conditioning on sex and either rs4779584 or rs58658771. The top results are shown in 3.2. A visual summary of the output after conditioning on rs4779584 is available in figure 3.5.

Conditioning on	SNP	Position (hg37)	P
rs4779584			
	rs17816465	33156386	7.10×10^{-7}
	rs16969681	32993111	2.03×10^{-5}
	rs16969344	32991713	6.18×10^{-5}
	rs112503426	33155404	0.00020
	rs28630996	32993860	0.00021
	rs28399071	32992658	0.00023
	rs7494781	32988138	0.00026
	rs11632715	33004247	0.00074
	rs12593365	33107990	0.00096
	rs55746350	33127480	0.00096
rs58658771			
	rs17816465	33156386	2.06×10^{-6}
	rs16969681	32993111	7.51×10^{-6}
	rs16969344	32991713	2.48×10^{-5}
	rs28630996	32993860	8.40×10^{-5}
	rs28399071	32992658	9.48×10^{-5}
	rs16969862	32994001	0.00023
	rs7494781	32988138	0.00034
	rs112503426	33155404	0.00050
	rs55746350	33127480	0.00114
	rs12593365	33107990	0.00120

Table 3.2: Top CRC-associated SNPs following conditioning on rs4779584 or rs58658771. Importantly, rs16969681's signal remains strong.

A key finding was that, once again, rs16969681 was partially independent of rs4779584, and was still the best candidate for functional follow-up in its immediate region. This finding was reported in Lewis et al. (2014), along with a raft of experimental evidence showing the functional consequences of the SNP. I had no part in the wet lab work reported in Lewis et al. (2014), but its conclusiveness is important for my argument, so I have provided an incomplete summary of a few results in figure 3.6.

In short, rs16969681 lies in a region of active chromatin in *GREM1*-expressing CRC cell lines, and acts as an allele-specific enhancer in luciferase assays. The risk-associated allele, "T", has the higher activity (figure 3.6A).

3.2.1.1 Discovery and validation of rs17816465

As well as confirming the genetic evidence for rs16969681, the conditional analysis highlighted a second SNP. rs17816465 was the top SNP after both conditional tests, its P value largely unaffected by the new covariates (in the original "naïve" test, $P_{\text{rs17816465}} = 1.47 \times 10^{-6}$, $I^2 = 1.1\%$). This P value was small enough to warrant follow-up.

To create a large validation case-control set for rs17816465, I either manually genotyped or received summary data from a variety of CRC panels. The validation sets are described in more detail in §2.6.3, and are summarised here in table 3.3.

Cohort	Cases	Controls	Total
CORGIbcd	562	910	1472
COIN	1950	2162	4112
Colombia	820	822	1642
CORECT	5100	4831	9931
Finnish	1172	8266	9438
GECCO	10,711	13,379	24,090
ICR	3227	3613	6840
Mexico	904	866	1770
All	24,446	34,849	59,295

Table 3.3: Summary of rs17816465 validation cohorts.

Non-overlapping discovery and validation set rs17816465 tests were combined by meta-analysis. The result is shown in figure 3.7. Some of the larger validation case-control series, such as GECCO and CORECT, are themselves meta-analyses of many

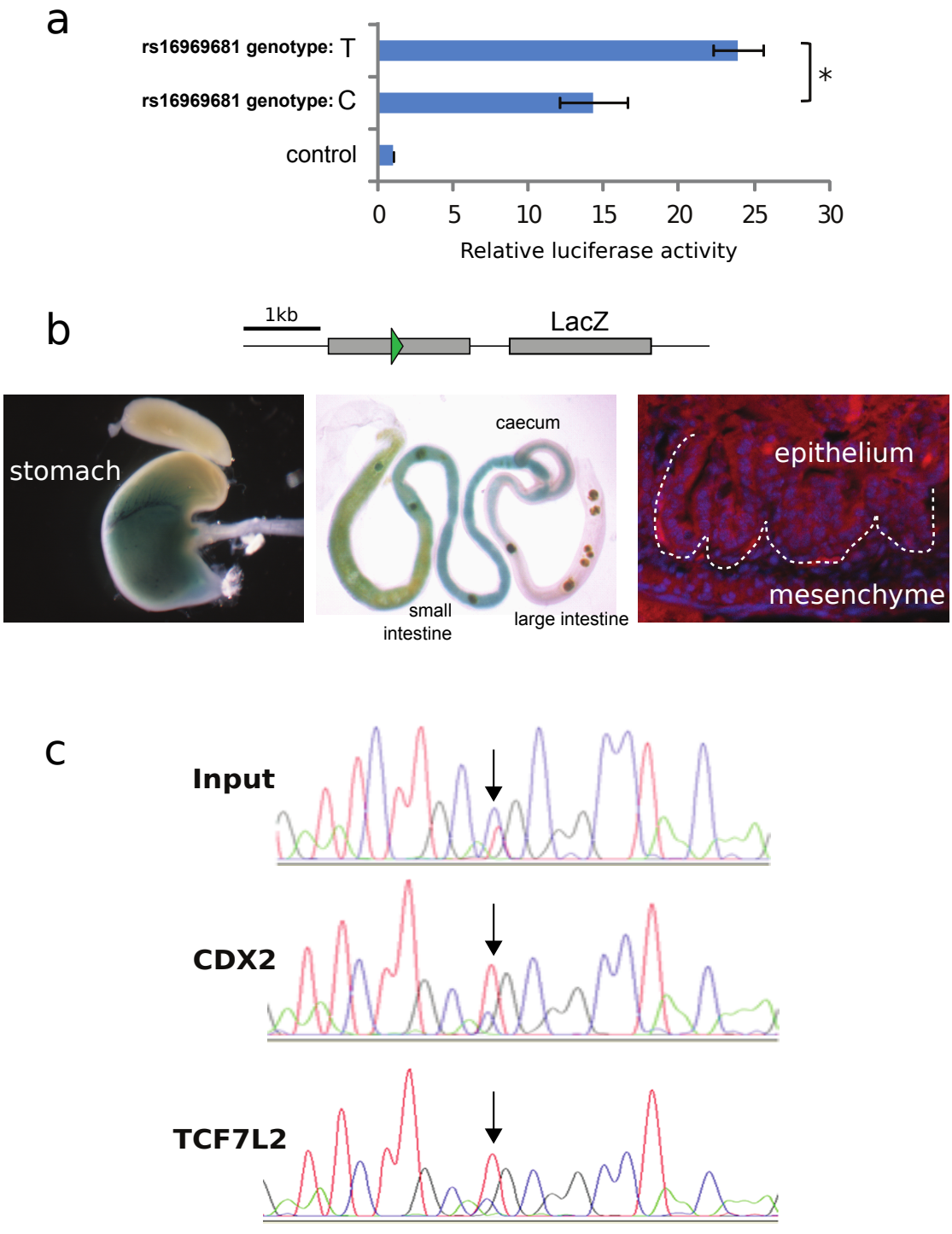


Figure 3.6: Highlight of evidence for rs16969681's biological role. *a*, the risk-associated allele has a significantly higher activity in luciferase enhancer assay; *b*, when placed upstream of LacZ reporter in embryonic mouse, rs16969681 drives expression (blue; red in confocal image) throughout the gut, including the epithelium; *c*, CDX2 or TCF7L2 ChIP is enriched for the risk-associated allele (T, red trace denoted by arrows) in heterozygous cells; see figure 6.9. All figures adapted from Lewis et al. (2014).

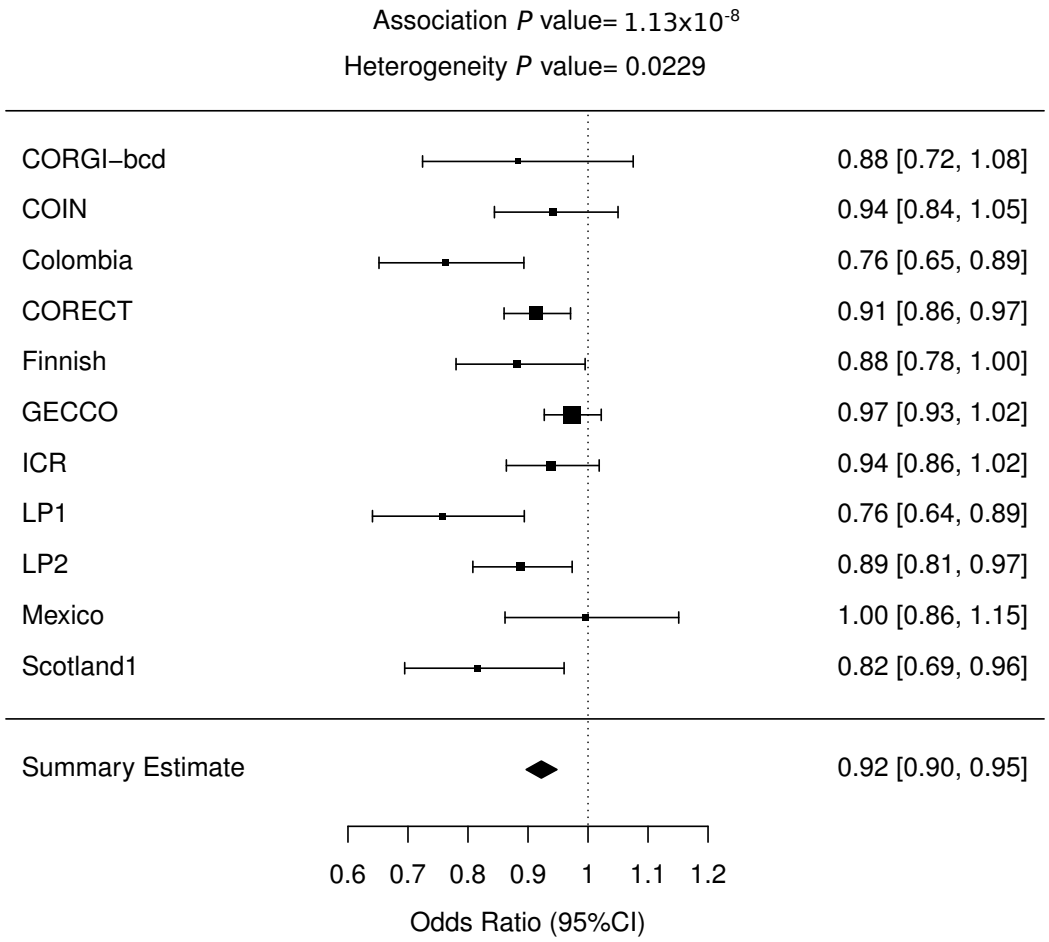


Figure 3.7: Forest plot of combined discovery and validation sets, showing genome-wide significant association of rs17816465 with CRC. The encoded allele is "G", the protective allele.

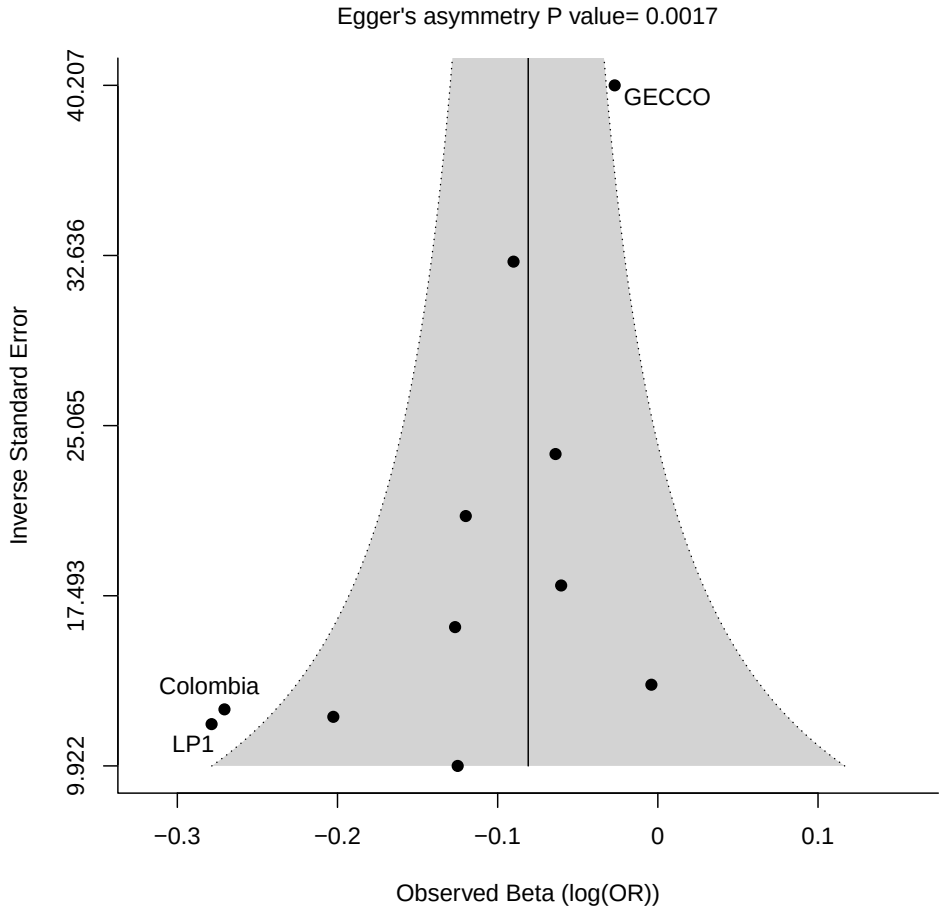


Figure 3.8: Funnel plot of combined discovery and validation sets, showing mild asymmetry diagnostic of inter-study heterogeneity. y-axis, $\frac{1}{se}$; x-axis, study β; grey area, expected 95% CI around the summary estimate (vertical line). Studies outside this interval are labeled.

smaller studies. It is perhaps unsurprising, therefore, that the combined analysis was significantly heterogeneous, with $I^2 = 51.8\%$. This heterogeneity of rs17816465 effect was not present in the discovery cohorts tested. An I^2 value $> 25\%$ indicates "moderate" study heterogeneity (Higgins et al., 2003), as can also be seen in the mild asymmetry of figure 3.8. Of note is that the largest study, GECCO, lies outside the 95% confidence interval corresponding to its expected inverse standard error in the absence of overall heterogeneity (figure 3.8, grey shade). Egger's regression test for funnel plot asymmetry returned a significant $P = 0.0017$, showing a systematic difference in reported effect size depending on study size. Together, these data suggest that the moderate I^2 is driven largely by GECCO; when GECCO is excluded from the meta-analysis $P_{\text{Egger}} > 0.1$.

However, for an exploratory validation test composed of studies performed with different genotyping technologies, $I^2 = 51.8\%$ is acceptable. Moreover, the meta-analysis confirmed rs17816465's risk signal, which became genome-wide significant ($P < 5 \times 10^{-8}$).

3.2.2 Further fine-mapping at *GREM1*

rs16969681 is an independent, experimentally validated CRC risk SNP. rs17816465 is genome-wide significant and largely unrelated to the tagging peak around rs4779584. But there is still risk association captured by rs4779584 that is unaccounted for by either SNP. Tomlinson et al. (2011) explained the residual signal by using stepwise regression to incorporate rs11632715 into their two-SNP model. I wished to advance my account of *GREM1*-adjacent risk using a similar approach.

The problem was to construct a multiple-SNP model that improved on rs4779584 or a correlated proxy while remaining biologically plausible, which is to say, without overfitting. In a way these criteria can only be judged subjectively, although the strength of different models can be measured quantitatively. Prior to stepwise model building, I made two assumptions. First, that rs16969681 deserved a place in any model, as its experimental evidence was strong. Although not in an eQTL for *GREM1* expression in either TCGA or published data (see figure 3.10), the SNP's open chromatin environment, allele-specific transcription factor binding, and nonspecific enhancement of LacZ in

mouse gut together constitute an unusually thorough account of the SNP's genome-wide significant CRC association. The lack of eQTL is not incompatible with the suggestion of Lewis et al. (2014): that the variant affects an enhancer only active in adenomas, not in normal or carcinoma tissue; a colorectal adenoma-only eQTL at rs16969681 has not been ruled out. It has better functional evidence than its correlates. Second, that parameter incorporation should be heavily penalised, on the intuition that the fewer SNPs included, the more likely the outcome to be biologically as well as statistically meaningful.

As such, I created a panel of SNPs by taking the top ($P < 10^{-6}$) remaining variants after conditioning on both rs16969681 and rs17816465. I read these "discovery" genotypes into R as continuous allelic dosages, along with rs16969681 and rs17816465. This fulfilled the first assumption, as it necessitated the inclusion of rs16969681 in any model built. I then performed bi-directional stepwise logistic regression, minimising, not on AIC, but on the "Bayesian" information criterion (BIC), which significantly more heavily penalises parameter number. This fulfilled the second assumption.

The output model was composed of cohort, rs16969681, rs17816465, and two other SNPs, both already discussed in §3.1.2: rs11632715 (from Tomlinson et al., 2011) and rs73376930 (from Whiffin et al., 2014). Together, these four SNPs produced the most parsimonious risk model as measured by BIC.

This output was encouraging, not only because both new SNPs were published genome-wide significant hits. In addition, the two SNPs were correlated: in both my cohorts and the 1000G CEU population, $D' = 0.87$ (see figure 3.9). What's more, they spanned the *GREM1* promoter CpG island, which lies at chr15:33,009,531–33,011,696 (Observed CpG/Expected CpG = 0.93). The positions of the four SNPs relative to *GREM1* and its CpG island are shown in figure 3.10. Together, these observations suggested that rs11632715 and rs73376930 were tagging a CRC-associated *GREM1* promoter haplotype.

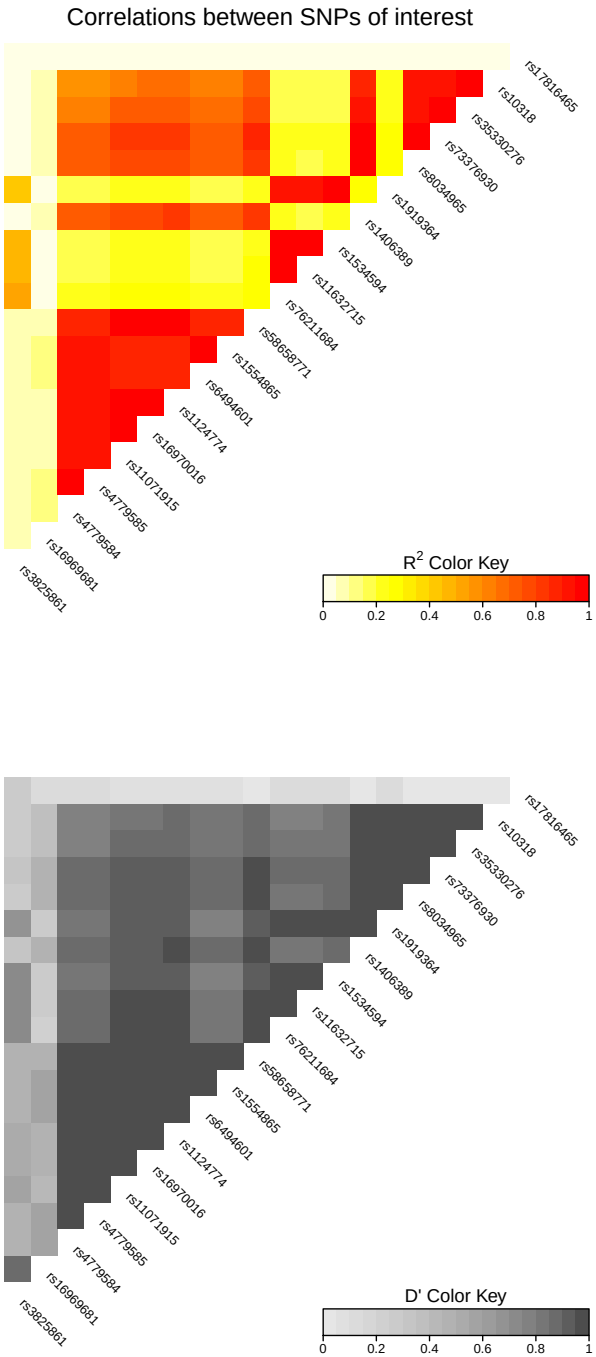


Figure 3.9: LD between the SNPs tested for model inclusion by stepwise regression, within the discovery cohorts of table 3.1. Top, r^2; bottom, D'.

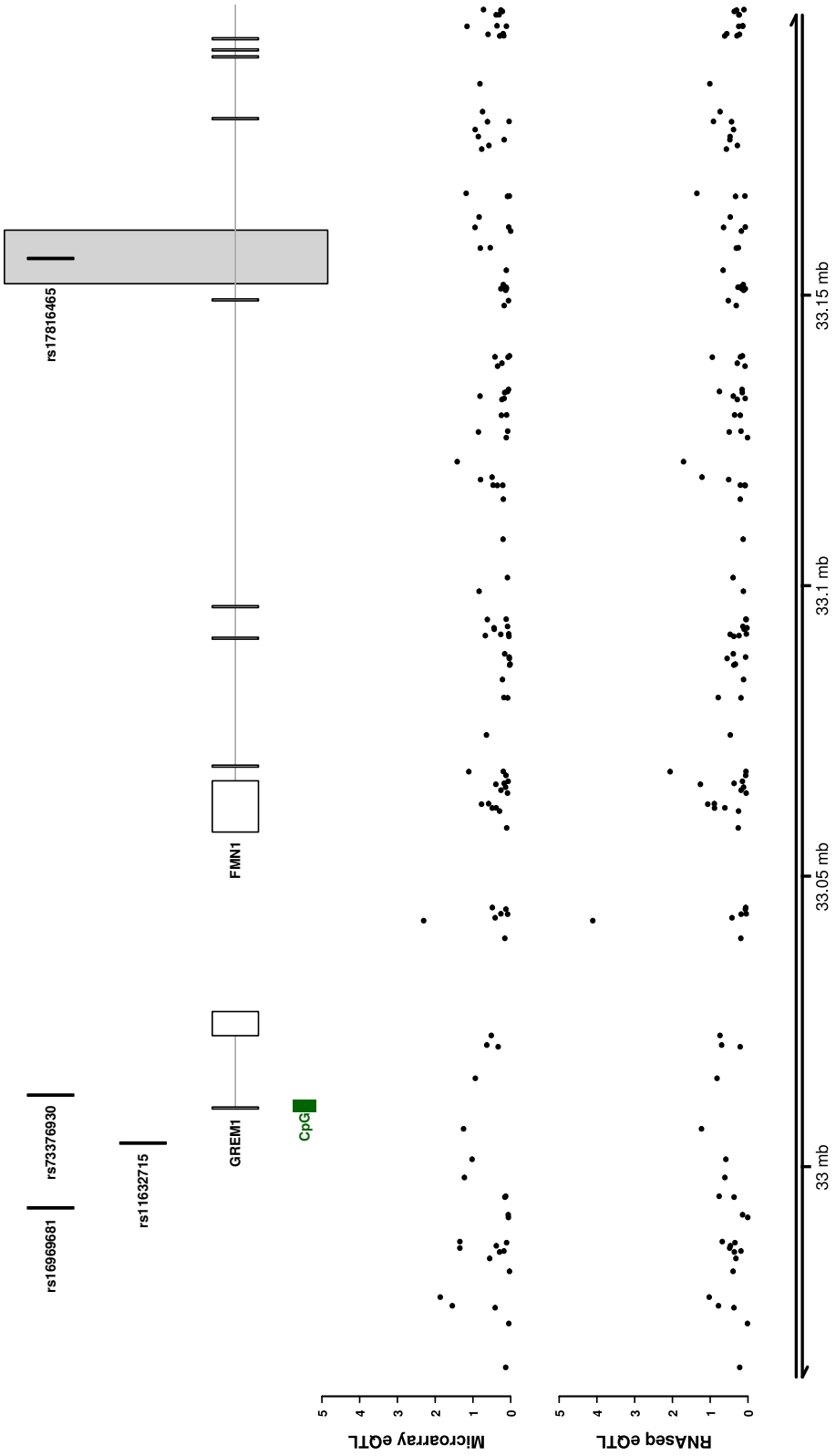


Figure 3.10: Map of chr15q13.3 showing the positions of a selection of CRC-associated SNPs, *GREM1* CpG island, and the absence of *GREM1* eQTL in TCGA cancer expression data, seen by microarray (top) or RNA sequencing (bottom). x-axis, coordinates on chromosome 15 (hg19); y-axes, negative log of the P value of correlation between copy number-conditioned expression value and genotype; white boxes, exons; green region, CpG island; grey region, span of ChIP interrogation.

3.2.3 Haplotypic analysis

I loaded the "best guess" (probability threshold 0.9) genotypes of the 20 SNPs in figure 3.9 into PLINKv1.07 for haplotype association testing. After conditioning on rs16969681 and rs17816465, an omnibus test of association at rs11632715-rs73376930 returned a significant $P_{\text{Omnibus}} = 3.15 \times 10^{-5}$, as expected. Testing the risk-risk haplotype rs11632715:A-rs73376930:G against all others gave $OR = 1.145$ at $P = 3.47 \times 10^{-6}$. Table 3.4 shows how obvious the haplotype frequency changes are between cases and controls.

rs11632715	rs73376930	$freq_{\text{controls}}$	$freq_{\text{cases}}$
A	G	0.168	0.197
G	G	0.013	0.0129
A	A	0.299	0.296
G	A	0.520	0.494

Table 3.4: Summary of haplotype frequencies in discovery cohort cases and controls.

There was no evidence for an independent effect of rs11632715 ($P = 0.29$) and only mild evidence for the independence of rs73376930 ($P = 0.0027$), driven by the great rarity of the G-G haplotype. Crucially, no omnibus association at the locus remained after controlling for the A-G signal ($P = 0.371$).

In other words, the evidence was promising for a moderately rare ($freq = 0.18$) promoter haplotype underpinning CRC risk at *GREM1*. I extended my analysis to include two candidate SNPs: one 53 bases upstream of the CpG island (rs1406389, from Whiffin et al., 2013), one (rs1919364) within the CpG island. Due to the high LD between these SNPs as visible in figure 3.9, this merely refined the definition of the haplotype to rs11632715:A-rs1406389:T-rs1919364:G-rs73376930:G, a block with frequency 0.183, $OR = 1.147$, and association $P = 1.26 \times 10^{-6}$ after conditioning on rs16969681 and rs17816465. For both additional SNPs, it was impossible to test for independence of effect, so perfectly were they correlated with rs73376930, but controlling for A-T-G-G left no remaining omnibus association signal ($P = 0.743$). In R, in a risk model containing sex, cohort, rs16969681, rs17816465, and rs1406389, all SNP covariates were significant at $P < 1 \times 10^{-4}$.

The most compelling explanation of these results is that there is a risk-associated haplotype at the *GREM1* promoter which has been tagged four times in the literature: by Tomlinson et al. (2011) at rs11632715, by Whiffin et al. (2013) at rs1406389, by Whiffin et al. (2014) at rs73376930, and by Du et al. (2016) at rs2293582. Of these four interlinked genome-wide significant CRC risk SNPs, probably rs1406389 is the best candidate for functional follow-up, as it lies closest to the promoter.

3.2.3.1 Epistatic analysis

Interaction tests are confounded by haplotypic effects and by high LD more generally (Wei et al., 2014). As such, the only sensible tests at *GREM1* were between rs17816465 and other candidates (where r^2 and D' are both considerably < 0.1). P values for this epistasis as calculated in PLINKv1.07 are shown in table 3.5; each but the rs17816465:rs11632715 interaction was significant after Bonferroni correction, though none comes close to the genome-wide interaction threshold of $P < 1 \times 10^{-12}$ (Becker et al., 2011).

SNP A	SNP B	$P_{\text{interaction}}$
rs17816465	rs11632715	0.0438
rs17816465	rs16969681	0.0109
rs17816465	rs1406389	0.0006
rs17816465	rs73376930	0.0005

Table 3.5: PLINKv1.07-generated P values for epistasis between model SNPs.

In dosage models in R, these interaction terms were sign-switched compared to their additive covariate components.

3.2.4 Bioinformatic function search: eQTLs

To search for eQTLs, I downloaded and queried tumour and normal TCGA expression data from colon and rectum samples, measured by both microarray and RNA sequencing. After regressing gene expression on average copy number within or spanning each gene's coordinates, I then regressed the unexplained variance in gene expression on the genotype at each SNP under an additive model. Full details are available in §2.7.12.

The underlying hypothesis of this work was that risk SNPs eventually affect gene

SNP	Gene	Tissue	Citation	r^2 with tag
rs4779584	*GREM1*	liver	Schadt et al. (2008)	1.0
rs1258734	*GREM1*	blood	Fehrmann et al. (2011)	0.019
rs727945	*FMN1*	lung	Hao et al. (2012)	0.296
rs3817591	*FMN1*	lung	Hao et al. (2012)	0.056
rs17816711	*FMN1*	lung	Hao et al. (2012)	0.00
rs17525764	*FMN1*	lung	Hao et al. (2012)	0.00

Table 3.6: Published eQTLs at the *GREM1* locus. LD is within 1000G relative to rs4779584.

expression, so they or their tags should lie in eQTLs. A tag in LD with an eQTL would implicate in disease risk both the locus and the gene whose expression varied.

At the *GREM1* locus I tested nearby genes for eQTLs in this way, including *GREM1* and *FMN1*, although *FMN1* had only RNAseq expression data available. No gene displayed a multiple testing-corrected significant correlation with any nearby genotype. Figure 3.10 illustrates this, showing flat P values for eQTL with *GREM1*.

However, eQTLs are tissue- and even cell type-specific. There are regional eQTLs at 15q13.3 in other studies on less relevant biological backgrounds. These are summarised in table 3.6.

3.2.5 Functional validation of rs17816465

rs17816465 is associated with CRC risk. But this could mean that it is a variant within a functional element, changing some process directly; or it could be a secondary tag, correlated with an untyped variant which is itself effecting the important biological change. In the absence of a better candidate SNP for functionality, I attempted a number of molecular assays to try to distinguish between these two possibilities. The majority of these experiments, however, were inconclusive.

3.2.5.1 ChIP

Certain histone modifications are correlated with — or possibly control; the issue is controversial — the activity of their surrounding DNA. In particular, the dimethylation of lysine 4 of histone H3 (H3K4me2) is a mark of transcription factor binding (Wang et al., 2014b), and H4 pan-acetylation (H4ac) marks active chromatin more generally, as the modification lessens the subunit's positive charge and therefore its hold on DNA.

To see whether rs17816465 was overlaid by either of these marks, I performed chromatin immunoprecipitation (ChIP), as described in §2.10.2, with anti-H3K4me2 and anti-H4ac antibodies. This asked whether rs17816465 lay at an enhancer element active in the cell lines I tested. These lines were chosen for their above-zero expression of *GREM1* by qRT-PCR (data not shown), and to span the possible genotypes of rs17816465, in case of a visible allele-specific difference. They are listed in table 3.7.

Cell line	rs17816465 genotype
HCD57	GG
LS180	GG
COLO320	GA
SW1222	GA
T84	AA

Table 3.7: Cell lines used for rs17816465 histone ChIP. All express *GREM1* as determined by qRT-PCR.

Each immunoprecipitation was performed in biological triplicate, and each replicate then underwent qPCR in technical triplicate. Analysis, normalising against both input and a negative control, was as detailed in §2.10.2. The results are shown in figures 3.11 and 3.12. Though messy, the enrichment of both marks is essentially flat across the 9213bp probed. No consistent pattern is shared by the cell lines with identical rs17816465 genotypes.

3.2.5.2 3C

Another common assay for SNP function is chromatin conformation capture (3C). This tests for the proximity of restriction fragments within the 3D environment of the nucleus; a positive result is said to indicate the "looping" together of linearly remote DNA segments. The EcoRI-HF restriction fragment surrounding *GREM1*'s transcription start site was my bait region. Figure 3.13 shows the degree of capture by cross-linking to this bait region of four restriction fragments around rs17816465, as well as of the two fragments immediately downstream of the *GREM1* promoter.

Each point in figure 3.13 is the normalised average of PCR triplicates for three biological replicates in the rs17816465 heterozygous cell line, COLO320. As expected, following crosslinking the regions immediately adjacent to the *GREM1* promoter were

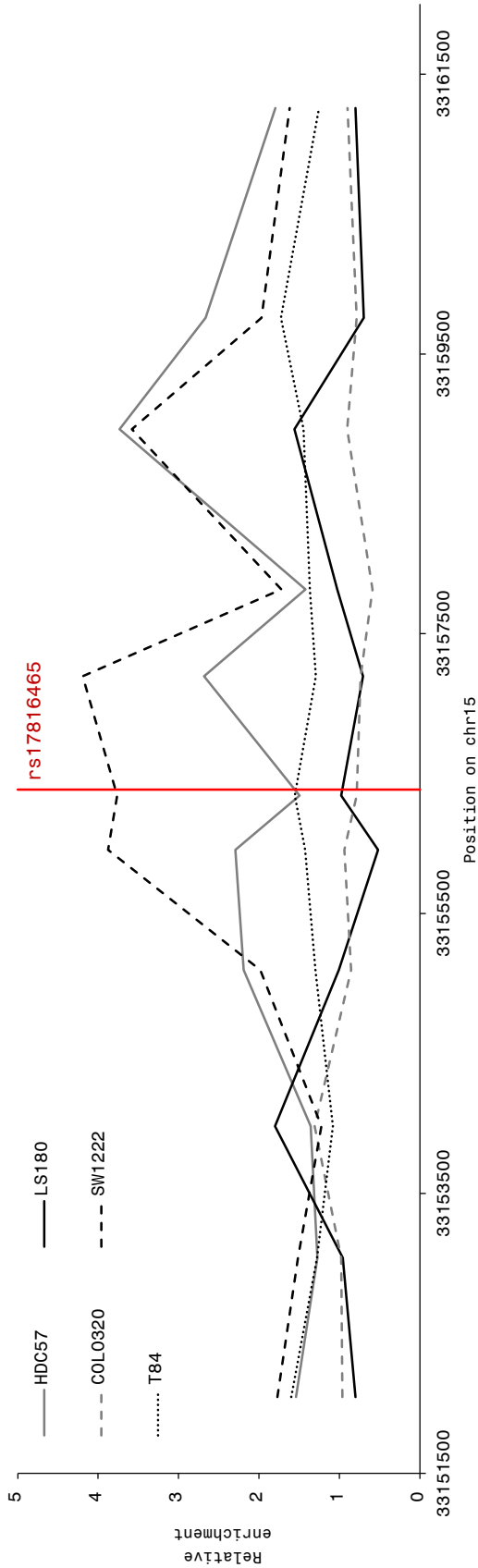


Figure 3.11: Anti-H4 pan-acetylation ChIP surrounding rs17816465.

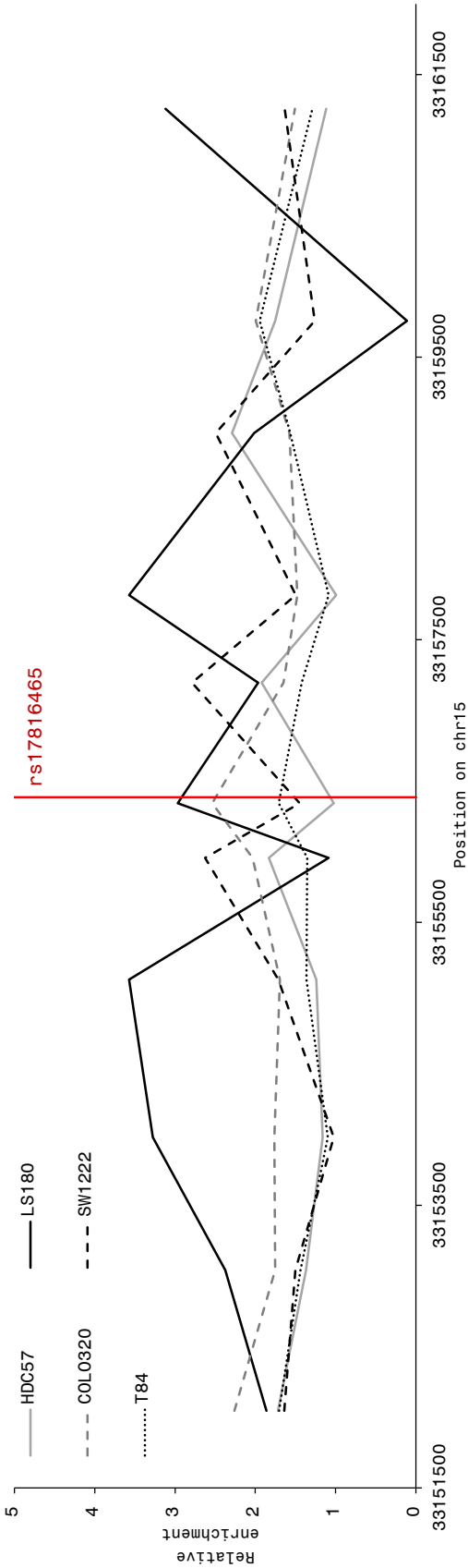


Figure 3.12: Anti-H3K4 dimethylation ChIP surrounding rs17816465. Line type indicates genotype; see table 3.7.

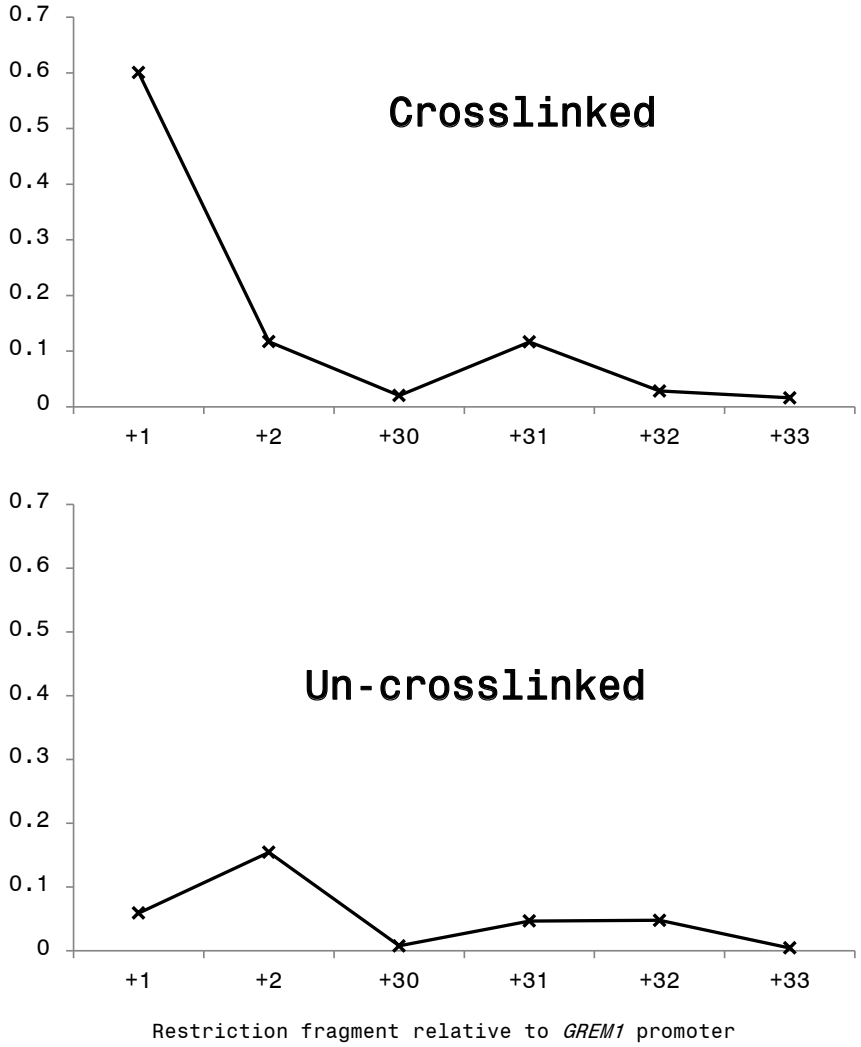


Figure 3.13: 3C assay searching for rs17816465 looping to *GREM1* in COLO320 cells. y-axis, relative enrichment. Top, little to no looping is seen following crosslinking. Bottom, the un-crosslinked negative control is also flat, even at the +1 fragment adjacent to the bait.

enriched and amplified. However, the *FMN1* intron containing rs17816465, far downstream, was not enriched (fragments +30–+33).

3.2.5.3 Allele-specific luciferase activity

The 2kb stretch of DNA surrounding rs17816465 was cloned by Dr. Barbara Fortini, a collaborator in Prof. Graham Casey's group, into the pGL3 promoter vector (Promega). The cell lines HCT-116, which does not express *GREM1*, and SW480, which expresses the gene at a low level, were transfected on at least three occasions with at least three separate clones, with the rs17816465 element inserted in both orientations. The results are shown in figure 3.14. In the reverse orientation only, the protective allele (G) of rs17816465 showed allele-specific enhancer activity in both cell lines ($P_{\text{HCT-116}} = 0.0024$; $P_{\text{SW480}} = 0.0025$). The risk-associated allele (A) enhancer activity was indistinguishable from control.

3.3 Discussion

I have reported a new CRC risk-associated variant at the *GREM1* locus, which lies far downstream of the gene, within a *FMN1* intron. This SNP does not loop to the *GREM1* promoter and does not lie in a region of regulatory chromatin in the cell lines and conditions that I interrogated. It does, however, display allele-specific luciferase activity.

There is additional cause to find the rs17816465 signal exciting. It is the *limb deformity* mouse. Although most *ld* alleles disrupt *Formin* rather than *Grem1*, Zúñiga et al. (2004) showed that the pathogenic change was always to disrupt *Grem1* expression, by deleting a control region within the *Formin* gene. This broad control region has three areas of strong conservation, which Zúñiga et al. (2012) called human-mouse-chicken-opossum conserved sequences 1–3 (HMCO1–3). The location of these sequences in the human genome is shown in figure 3.15b.

Zúñiga et al. (2012) focused in particular on HMCO1 in mice, whose presence is necessary and sufficient for *Grem1* expression in posterior limb bud mesenchyme. HMCO1 is bound by GLI3 and is also downstream of BMP signal. It integrates the many sig-

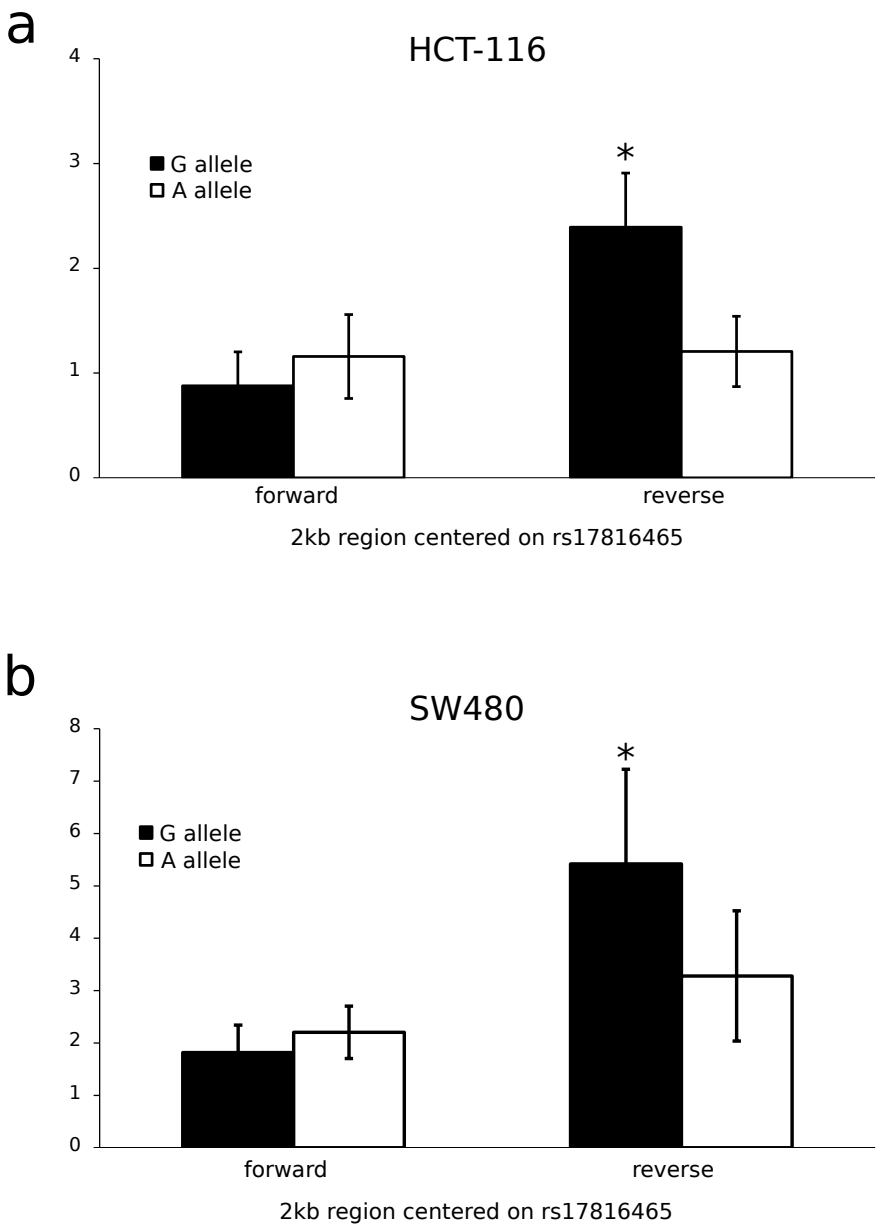


Figure 3.14: rs17816465 Allele-specific luciferase assay changes. *a*, HCT-116 cells; *b*, SW480 cells. Error bars represent standard deviation of the mean of four biological replicates, each performed in triplicate.

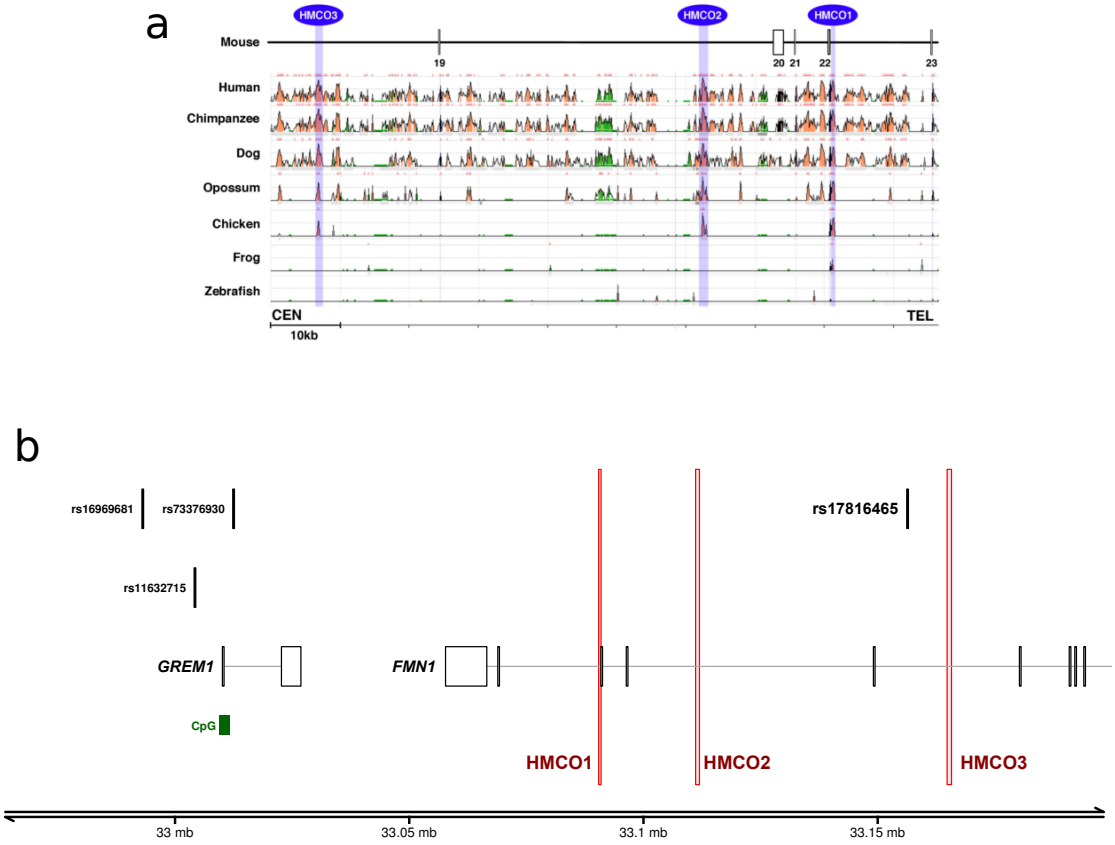


Figure 3.15: Conserved regions at the *GREM1* locus. *a*, Zúñiga et al. (2012) figure highlighting conserved regions within mouse *Fmn1*; *b*, the same regions overlaid on my simplified map of the human *GREM1* locus. Note that rs17816465 lies within the same intron as HMCO3.

nalling inputs in the developing limb bud. HMCO2 and 3 were not necessary for *Grem1* expression in this milieu, but could be important in other tissues or developmental stages. rs17816465 is close to HMCO3.

The mixed molecular evidence this thesis provides does not conclusively confirm or deny the function of rs17816465. My hypothesis for its risk signal remains that it is a looping regulatory element, possibly related to HMCO3, that influences *GREM1* expression in *cis*. The activity of the SNP's locus in a luciferase enhancer assay is midly in favour of this model, as is its significant sign-switching epistasis with other *GREM1* SNPs, which could be due to the competition of multiple risk-associated regulatory elements impeding each others' effects on the gene. However, the 3C and histone marker ChIP at rs17816465 described make this less likely. A more conclusive test of the activity of rs17816465 would be to use CRISPR-Cas9 to selectively edit heterozygous *GREM1*-expressing CRC cell lines, adding or removing the risk-associated allele on identical genetic backgrounds (as in Liu et al., 2017). Such cell lines would minimise the confounding differences present in my work, providing superior controls for expression analysis and histone and transcription factor ChIP. If this were successful, combinations of CRISPR-edits could further assess the possibility of competetion between regulatory GWAS elements at *GREM1*.

As well as providing another CRC risk SNP at a locus that one might reasonably have held to be saturated with information, the case of rs17816465 is particularly interesting as an illustration of the challenges and pitfalls of post-GWAS research. All loci will be different, and I shall try to avoid over-extrapolation, but my work does highlight three general principles.

The first is the tension between variant-based and model-based accounts of risk association. Lewis et al. (2014) and the work reported above both show that by not assuming a single causative variant at a GWAS locus, a richer biological account can arise. Splitting a reported signal into multiple SNPs can be the right thing to do, even though no individual SNP in that model might have as low a P value of association as the original tag. Bioinformatic tools to assess the likely number of causal SNPs at an association signal locus are underdeveloped (Su et al., 2009). Once a locus is identified

and validated, I'd argue that *AIC* is a better test of a model than a parameter's *P* value. The counterargument to this, which is quite legitimate, is that split signals become overfitted.

However, the case for rs4779584 capturing multiple causal SNPs is no longer merely statistical: rs16969681 has relevant biological effects (Lewis et al., 2014). And even if not accepted, the split signal at *GREM1*, first reported by Tomlinson et al. (2011), could at least have been entertained by subsequent papers. Instead, it's never mentioned. A non-exhaustive list of later papers that only consider rs4779584 at the locus includes: Whiffin et al. (2013), Whiffin et al. (2014), Zhang et al. (2014a), Timofeeva et al. (2015), Baert-Desurmont et al. (2015), Chen and Tian (2016), and Abulí et al. (2016). Conversely, Kupfer et al. (2014) corroborated separate independent signals at *GREM1* in an African American case-control cohort, although they did not successfully impute rs16969681.

Second, serial meta-analyses of many small cohorts can create unnecessarily heterogeneous signal. My validation of the CRC association of variation at rs17816465 had a significant $I^2 = 51.8\%$. Larger, more consistently recruited data sets are preferable. For this reason, the imminent delivery of cancer patient data from the UK National Health Service's 100,000 genomes project is welcome (Siva, 2015).

The third principle that my work on rs17816465 illustrates is the frustration of molecular biology in a post-GWAS context. The lack of strong functional evidence at rs17816465 is disappointing, but compatible with the SNP's activity in untested tissue types or under different stimuli. Individual assays are insufficient to test every CRC GWAS candidate. What's required is more high-throughput study designs, such as the "proteome-wide analysis of SNPs" proposed by Butter et al. (2012) and used at CRC GWAS loci by Liu et al. (2017). This looks for allele-specific trascription factor binding changes to DNA elements by subjecting those elements to normal and $^2H_4$-lysine-labeled proteins, which can be distinguished by mass spectrometry. Although Liu et al. (2017) focused on one interaction identified in this way, between rs1800734 and TFAP4 at *MLH1*, they also reported 730 other allele-specific transcription factor binding patterns at 116 CRC GWAS-related SNPs.

This sort of assay may help to explain the haplotypic signal in my model of *GREM1*

locus risk. I have reported the inclusion of this component, represented by rs11632715 and rs73376930, due to its minimisation of *AIC* in stepwise regression. I hypothesise that a (presumably single) variant lying on this haplotype is associated with CRC risk. This variant may have been typed already, or may be an untyped indel or SNP. There are multiple correlated candidates for function, including rs1534594, rs1406389, and rs72360084. These SNPs are conserved, exhibit promoter histone marks in gastrointestinal tissue, change transcription factor binding sites, and are reported eQTLs in Haploreg v4.1 (Ward and Kellis, 2012). A combination of GWAS fine-mapping in other populations — whose distinct LD structure may rule out certain candidates — and high-throughput post-GWAS assays should be performed to further clarify the *GREM1* promoter variation that influences CRC formation.

Chapter 4

Germline Variants May Influence Colorectal Cancer Driver Mutations

4.1 Introduction

In this chapter, I continue the fine-mapping of chapter 3, moving to the *POLD3* locus. I then explore whether the current collection of CRC GWAS risk SNPs might influence cancer subtype or behaviour, beyond merely increasing risk of incidence.

4.1.1 *POLD3*

Variants near *POLD3* were first associated with CRC by Dunlop et al. (2012). A meta-analysis of five of the seven cohorts used in this thesis was followed by validation in a number of smaller sets, including one from Japan. The combined association tests implicated in CRC risk a SNP in a *POLD3* intron, rs3824999 ($P = 3.54 \times 10^{-10}$). This signal has been replicated in Asian cohorts (Zhang et al., 2014a). However, Dunlop et al. (2012) also fine-mapped the locus and found an improved signal at rs72977282, upstream of the gene. It's odd, and possibly bespeaks a lingering suspicion of imputation, that rs3824999 is always described as the tag SNP at *POLD3* despite the reporting of a superior tag in the same paper.

The locus was multiply revisited, like *GREM1*. Whiffin et al. (2013) declared an improved association at rs11604752 or rs11236161, depending on whether the CGI set was combined in their imputation reference panel. These SNPs are reasonably distant from the original hit: rs11604752 and rs11236161 are 68kb and nearly 54kb centromeric

to rs3824999, respectively. Whiffin et al. (2014) then reported an indel 42kb upstream of the gene, before Du et al. (2016) returned to the first result, naming rs72977282 as their top SNP.

POLD3 is a subunit of DNA polymerase δ (Polδ), which synthesises DNA during replication. Until recently it was thought that Polδ only acted on the lagging strand, leaving leading strand synthesis to DNA polymerase ϵ (Pursell et al., 2007). This was largely based on evidence from a Polδ-mutant yeast strain showing a preponderance of lagging strand errors (McElhinny et al., 2008). However, in the absence of leading strand repair processes this selectivity disappears, and it now seems that Poldδ is the primary replicative enzyme, with Polϵ supplementing and correcting on the leading strand (Johnson et al., 2015). In any case, POLD3 is not the catalytic subunit of Polδ, but rather links the complex to a sliding clamp protein, PCNA. This interaction increases the processivity of the polymerase (Rahmeh et al., 2012).

4.1.2 Polygenic risk scores

Individual GWAS SNP effects on disease incidence and outcome risk may be too small to detect. It is therefore sensible to combine individual SNPs into overall risk scores, most often by summing alleles weighted by their published effect sizes. These scores have exhibited some clinical utility in breast cancer, where almost 100 risk alleles have been uncovered by GWAS (Michailidou et al., 2015).

Mavaddat et al. (2015) constructed a weighted score of 77 SNPs, which was significantly higher in 33,673 breast cancer cases than in 33,381 controls. Individuals with the highest 1% of risk scores were about three times as likely to get breast cancer as the middle quintile. In a smaller case set of women given tamoxifen prevention therapy, the 88-SNP genetic risk score of Cuzick et al. (2017) was significantly associated with breast cancer risk, especially with oestrogen receptor-positive cancer. More importantly, the score in combination with the women's Tyrer-Cuzick risk factor (incorporating *BRCA* status, family history, and environmental risk factors; see Tyrer et al., 2004) was more predictive than either method alone. This suggests genuine clinical utility.

Breast cancer risk is different for oestrogen receptor-positive and -negative tumours

(Kuchenbaecker et al., 2014), which confounds attempts to determine whether the genetic architecture of *BRCA1/2* mutation carriers is also distinct from more sporadic cancer (Gaudet et al., 2013). In these breast cancer subsets, polygenic risk scores remain predictive. GWAS risk scores accurately describe disease risk in both women and men with germline *BRCA1* and *BRCA2* mutations (Kuchenbaecker et al., 2017; Lecarpentier et al., 2017). Since these individuals are predisposed to cancer, the measurement of their common risk input provided by polygenic risk scores is additionally useful.

4.1.3 Association with disease outcome

Traditional GWAS looks for association between common variation and disease incidence. Naturally, it's also possible to correspond genotype to other variables, such as cancer subtype, chemotherapy response, or survival time.

Some recent reports have found survival associations in CRC. In a small retrospective study, Liu et al. (2015b) saw decreased disease-free survival in patients associated with the major allele of rs2335052, independent of initial CRC stage. The SNP was not in an eQTL for its nearest gene, *GATA2*. Morris et al. (2015) associated the minor allele of the *BMP4* GWAS SNP rs4444235 with a worse prognosis in 4327 CRC patients, correcting for age, sex, Dukes stage, tumour site, and year of diagnosis. However, a number of other reasonably well-powered studies found no multiple-testing corrected survival associations with common variation, whether or not this variation had previously been highlighted by GWAS (Abulí et al., 2013; Dai et al., 2012; Hoskins et al., 2012; Huhn et al., 2014).

One variant's downstream biological effects, including its impact on cancer suvival, have been extensively studied. This is the *MDM2* promoter SNP rs2279744. The SNP was first reported as a sarcoma and breast cancer risk factor, accelerating cancer formation in both the general population and in those with germline *TP53* mutations, who suffer Li-Fraumeni syndrome (Bond et al., 2004). The G allele at rs2279744 also hastens CRC formation in women (Bond et al., 2006b; Menin et al., 2006) across ethnicities (Wang et al., 2014b; Wo et al., 2011). The female-only effect is due to the SNP's interaction with oestrogen. The interaction is clearest in breast cancer, where rs2279744 was only associated with earlier intraductal carcinoma in women with high expression

of oestrogen receptor. This relationship was stronger in pre-menopausal women (Bond et al., 2006a).

MDM2 negatively regulates p53. A missense variant in *TP53*, rs1042522, interacts with rs2279744 to increase cancer risk across tumour types (Wan et al., 2011). This interaction with p53 extends to the variant's effects on breast cancer survival. Boersma et al. (2006) discovered a significant interaction term between rs2279744 genotype and *TP53* expression for breast cancer survival. And a Finnish group found that another gene in the p53 network, *PRKAG2*, contained a variant which improved breast cancer survival only in the context of rs2279744 minor allele presence (Jamshidi et al., 2013). This work illustrates the fine resolution achievable in the exploration of germline effects on cancer biology.

One additional variable relatable to germline variation is somatic mutation spectrum. SNPs might enourage certain oncogenic mutations over others. This possibility was recently explored by Carter et al. (2017) in nearly 7000 TCGA patients. The authors examined 138 genes important in cancer; seventeen germline-somatic correlations were validated. For example, variation at 19p13.3 was associated with a large increase in *PTEN* somatic mutation. Oddly, some classical cancer genes, including *TP53*, did not display a similar germline relationship.

4.2 Results

4.2.1 Fine-mapping risk near *POLD3*

I imputed genotypes at the chromosome 11 CRC risk locus as at the *GREM1* locus. The region between chr11:73,553,000–75,103,999 was imputed in the seven case-control cohorts available against a combined reference panel of 1000G and CGI.

After filtering on info score, the top SNP at the *POLD1* locus remained rs72977282, as shown in figure 4.1. As would be expected given the similarity of cohorts to Dunlop et al. (2012), the SNP remained genome-wide significant at $P = 5.90 \times 10^{-9}$. The association signal collapsed after conditioning on rs72977282, including at the original tag rs3824999 (figure 4.2). When modeled in R as allele dosage, a single-SNP effect

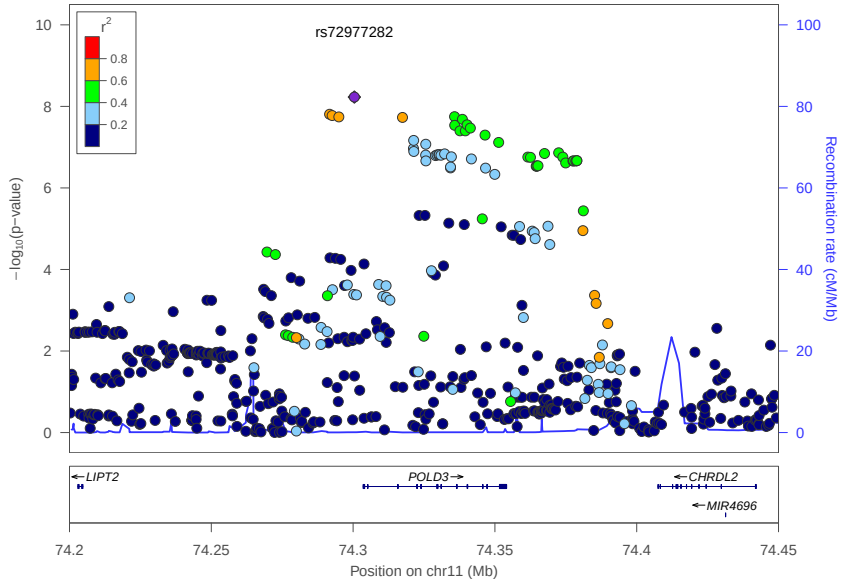


Figure 4.1: Unconditional meta-analysed P values at the *POLD3* locus. The top SNP, rs72977282, is coloured in purple and is the reference for LD values.

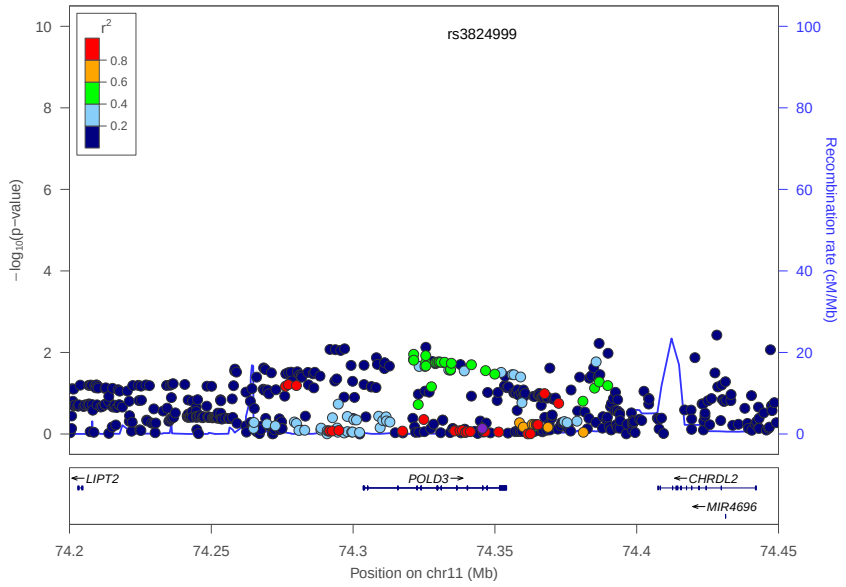


Figure 4.2: P values at the *POLD3* locus after conditioning on rs72977282, showing no remaining signal. The original tag SNP, rs3824999, is coloured in purple and is the reference for LD values.

with sex and cohort as covariates had an almost exactly equivalent model fit ($AIC = 27191.72$) to a series of two-SNP models (for example rs72977282+rs3824999, where $AIC = 27190.64$), but the additional covariate abolished the significance of both terms: they were not independent signals. This was entirely in keeping with the very strong LD between the top SNPs at the locus, which appeared moderate by r^2 (figure 4.3), but which was almost absolute when measured by D' (figure 4.4).

This LD structure was not unique to my cohorts. The top 56 SNPs were also in almost total LD in 1000G. The strong possibility existed, then, that there was a haplotypic signal at *POLD3*.

As at *GREM1*, I loaded "best guess" genotypes of the 56 SNPs with $P < 10^{-5}$ into PLINKv1.07. An omnibus test of haplotypic association at rs72977282-rs3824999 gave $P_{\text{Omnibus}} = 7.6 \times 10^{-4}$, with rs72977282:T-rs3824999:T showing a protective effect at $P = 1.76 \times 10^{-4}$ ($OR = 0.924$). The top SNP, rs72977282, did not have an effect independent of this haplotype ($P = 0.615$); nor was there remaining signal after conditioning on the haplotype ($P_{\text{Omnibus}} = 0.589$). Table 4.1 illustrates this simple block, where the risk signal comes from rs72977282's minor allele and rs3824999's major allele.

rs72977282	rs3824999	$freq_{\text{controls}}$	$freq_{\text{cases}}$
T	T	0.517	0.497
A	G	0.411	0.430
T	G	0.0711	0.0727

Table 4.1: Summary of *POLD3* haplotype frequencies in discovery cohort cases and controls. The A-T haplotype was below PLINK's frequency theshold of 0.001.

Unlike at *GREM1*, however, it was unclear where the boundaries of the associated haplotype lie. A logistic regression on the genotypes of the 56 top "plateau" SNPs defined a block incorporating every SNP ($freq = 0.401$) associated with CRC risk at $P = 7.53 \times 10^{-5}$; if instead a sliding window tested 15-SNP haplotypes for a joint effect, P_{Omnibus} was significant for every window. This was largely a result of the high LD between the SNPs: all risk association lay in a large block.

I constructed a preliminary delineation of the risk haplotype by providing PLINKv1.9 with best-guess genotypes of the imputed region, estimating haplotypes using the definition of Gabriel et al. (2002). SNPs as far as 10,000kb apart were considered for coin-

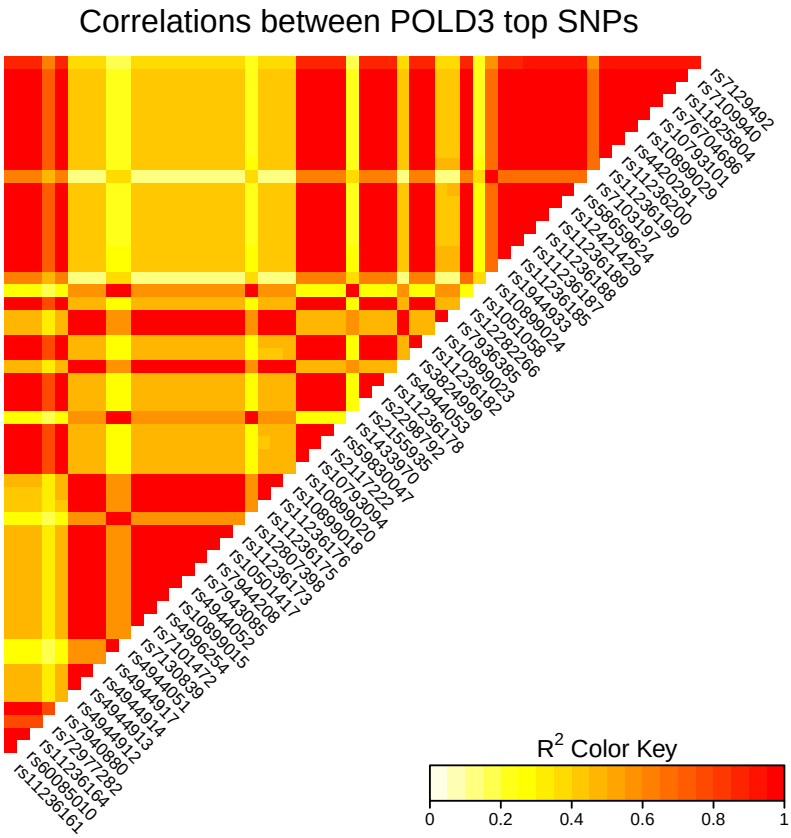


Figure 4.3: Correlation between all SNPs at the *POLD3* locus with $P < 10^{-5}$, within all cohorts, by r^2.

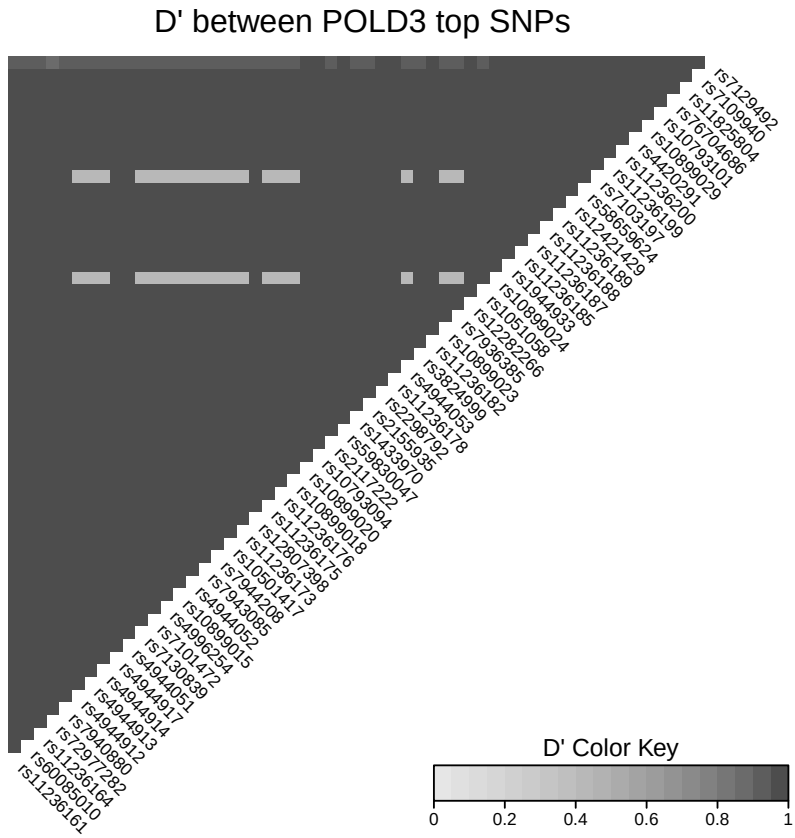


Figure 4.4: Correlation between all SNPs at the *POLD3* locus with $P < 10^{-5}$, within all cohorts, by D'. The LD is almost total, suggesting a single risk haplotype.

heritance and a lenient within-block informative fraction cutoff of 0.5 was used. These relaxed criteria provided the boundaries of a broad block, stretching from chr11:74,303,841–74,381,181. This area almost exactly covers the *POLD3* gene.

GWAS association is unusually uninformative at *POLD3*. rs72977282 tags a large risk haplotype that spans the gene and its surroundings.

4.2.2 Bioinformatic function search: eQTLs

As at the *GREM1* locus, I searched the 11q13.4 locus for eQTLs within TCGA colon and rectum cancer samples. No significant association with *POLD3* expression existed in RNAseq or microarray data (figure 4.5). I did see one genome-wide significant RNAseq eQTL with *CHRDL2*, at rs12284314 ($P = 4.29 \times 10^{-8}$). But this SNP is over a megabase away from the GWAS signal, and is not correlated with it.

There are three published eQTLs at the region: between rs4944950 and *CHRDL2* in liver (Schadt et al., 2008), and between *POLD3* and rs6592574 in lymphoblastoid cells (Fehrmann et al., 2011) and rs4944933 in liver (Innocenti et al., 2011). No SNPs correlated with rs72977282 at $r^2 > 0.6$ were in eQTLs reported by Haploreg v4.1 (Ward and Kellis, 2012).

4.2.3 The impact of risk SNPs on cancer biology

It's typical to end an association study with open-ended lists of candidate functional variants correlated with tag SNPs, awaiting molecular validation as in chapter 3 or Lewis et al. (2014). I wondered whether it was also possible to look for the biological effects of risk-associated SNPs *in silico*.

To some extent, I had already tested this idea by searching the GWAS loci for eQTLs, finding nothing. However, the TCGA is an imperfect resource for this search, being largely composed of cancers. I looked at the relationship between GWAS SNP genotype and other clinical variables available in public or in-house databases.

One such variable is recurrence-free survival time in CRC patients. High *GREM1* expression is associated with a shorter recurrence-free interval in an international cohort of all stages ($n = 403$; log-rank $P = 0.04$, figure 4.6a), and in an American stage II set

SNP	VICTOR OS	QUASAR2 OS	VICTOR RFS	QUASAR2 RFS
rs17816465	0.76	0.38	0.14	0.34
rs1406389	0.94	0.21	0.99	0.38
rs16969681	0.13	0.53	0.17	0.84
rs72977282	0.10	0.80	0.41	0.73

Table 4.2: P values of survival tests of individual SNP scores ($\beta\times$risk dosage). No individual test was significant. OS, overall survival; RFS, recurrence-free survival.

($N = 90$; log-rank $P = 0.016$, figure 4.6b). *GREM1* expression is also significantly lower in stage I CRC than in all other stages (ANOVA $P = 0.00005$; figure 4.6c). *POLD3* expression, however, does not discriminate recurrence-free survival time ($P = 0.26$ and 0.65 in each cohort, respectively).

The public databases used for figure 4.6 do not contain genotype information. For this, I had to turn to genome-wide imputation of the VICTOR and QUASAR2 trial samples. In Cox proportional hazards models of overall or recurrence-free survival, with age, stage and sex as covariates, no GWAS risk SNP from this thesis was significantly associated with altered survival time (table 4.2).

4.2.4 Overall GWAS Risk Score

The effect of individual SNPs on cancer outcome was imperceptible. I wondered whether an aggregate risk score composed of all known common risk modulators might show some discriminatory power. There's no reason to suspect this in particular: GWAS variants were discovered because they are associated with CRC incidence, not progression, and their effects might end at tumourigenesis. Nor is it implausible, however, that their subtle effects on colorectal biology might persist through malignant transformation.

I compiled a list of all published independent CRC risk SNPs whose association signal was significant in Europeans. This list is shown in table 4.3. It omits the chromosome X SNP, rs5934683 (Dunlop et al., 2012), because the VQ58 genome-wide imputation from which I constructed my score was autosomal only. Other reported European CRC variants not included in the score are explained in table 4.4.

The risk score was created by summing the risk allele dose at each locus, weighted by reported effect size (β) for every SNP. This score was normally distributed (see split

chr	Position	rsID	Citation
1	11856378	rs1801133	Hubner and Houlston (2007)
1	22587728	rs72647484	Al-Tassan et al. (2015)
1	183081194	rs10911251	Whiffin et al. (2014)
1	222045446	rs6691170	Houlston et al. (2010)
1	222164948	rs6687758	Houlston et al. (2010)
3	40924962	rs35360328	Schumacher et al. (2015)
3	66442435	rs812481	Schumacher et al. (2015)
3	169492101	rs10936599	Houlston et al. (2010)
5	1286516	rs2736100	Kinnersley et al. (2012)
6	36622900	rs1321311	Dunlop et al. (2012)
6	117822993	rs4946260	Schumacher et al. (2015)
8	117630683	rs16892766	Tomlinson et al. (2008)
10	8701219	rs10795668	Tomlinson et al. (2008)
10	14280702	rs12241008	Wang et al. (2014b)
10	16997266	rs10904849	Al-Tassan et al. (2015)
10	101345366	rs1035209	Whiffin et al. (2014)
10	101351704	rs11190164	Schumacher et al. (2015)
11	74300441	rs72977282	Dunlop et al. (2012) and this thesis
11	111171709	rs3802842	Tenesa et al. (2008)
12	4388271	rs3217810	Whiffin et al. (2014)
12	50539419	rs7138945	Spain et al. (2012)
12	50880216	rs7136702	Houlston et al. (2010)
12	51203371	rs1129406	Timofeeva et al. (2015)
12	111884608	rs3184504	Timofeeva et al. (2015)
12	117747590	rs73208120	Schumacher et al. (2015)
14	54410919	rs4444235	Houlston et al. (2008)
14	54560018	rs1957636	Tomlinson et al. (2011)
15	32993111	rs16969681	Tomlinson et al. (2011)
15	33004247	rs1406389	Whiffin et al. (2013) and this thesis
15	33156386	rs17816465	This thesis
16	32994756	rs9929218	Houlston et al. (2008)
16	86695720	rs16941835	Al-Tassan et al. (2015)
18	46453463	rs4939827	Broderick et al. (2007)
19	33532300	rs10411210	Houlston et al. (2008)
20	6404281	rs961253	Houlston et al. (2008)
20	6699595	rs4813802	Tomlinson et al. (2011)
20	47340117	rs6066825	Schumacher et al. (2015)
20	60921044	rs4925386	Houlston et al. (2010)

Table 4.3: The 38 SNPs used to contruct a GWAS risk score. All are significantly associated with CRC in Europeans and all represent plausibly independent signals. The X chromosome risk SNP rs5934683 from Dunlop et al. (2012) is not included as genome-wide imputation was autosomal only. Compare to table 1.4.

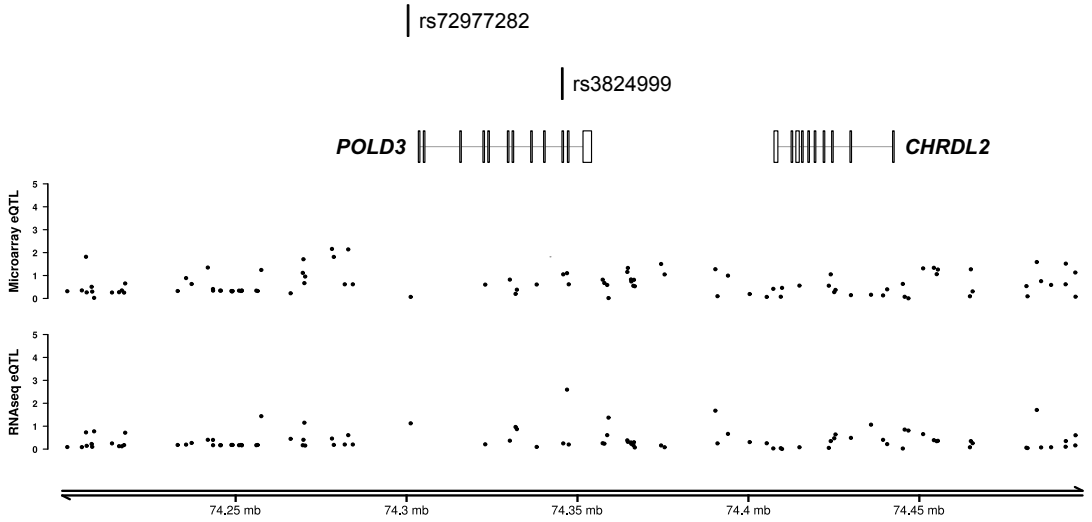


Figure 4.5: Absence of *POLD3* eQTL around *POLD3* in TCGA cancer expression data, seen by microarray (top) or RNA sequencing (bottom). x-axis, coordinates on chromosome 11 (hg19); y-axes, negative log of the P value of correlation between copy number-conditioned expression value and genotype. Important SNPs are indicated; not all regional genes are shown.

rsID	Citation	Reason
rs16888728	Timofeeva et al. (2015)	Not independent of rs16892766
rs6983267	Tomlinson et al. (2008)	Not independent of rs16892766
rs3824999	Dunlop et al. (2012)	Not independent of rs72977282
rs6580742	Timofeeva et al. (2015)	Reported but not genome-wide significant
rs12303082	Timofeeva et al. (2015)	Reported but not genome-wide significant
rs11169552	Houlston et al. (2010)	Not independent of rs1129406
rs4779584	Jaeger et al. (2008)	Not independent of rs16969681/rs1406389

Table 4.4: Published European CRC risk SNPs not included in the score, with justifications. Nor were the many Asian-only SNPs included, since the clinical cohort is entirely European.

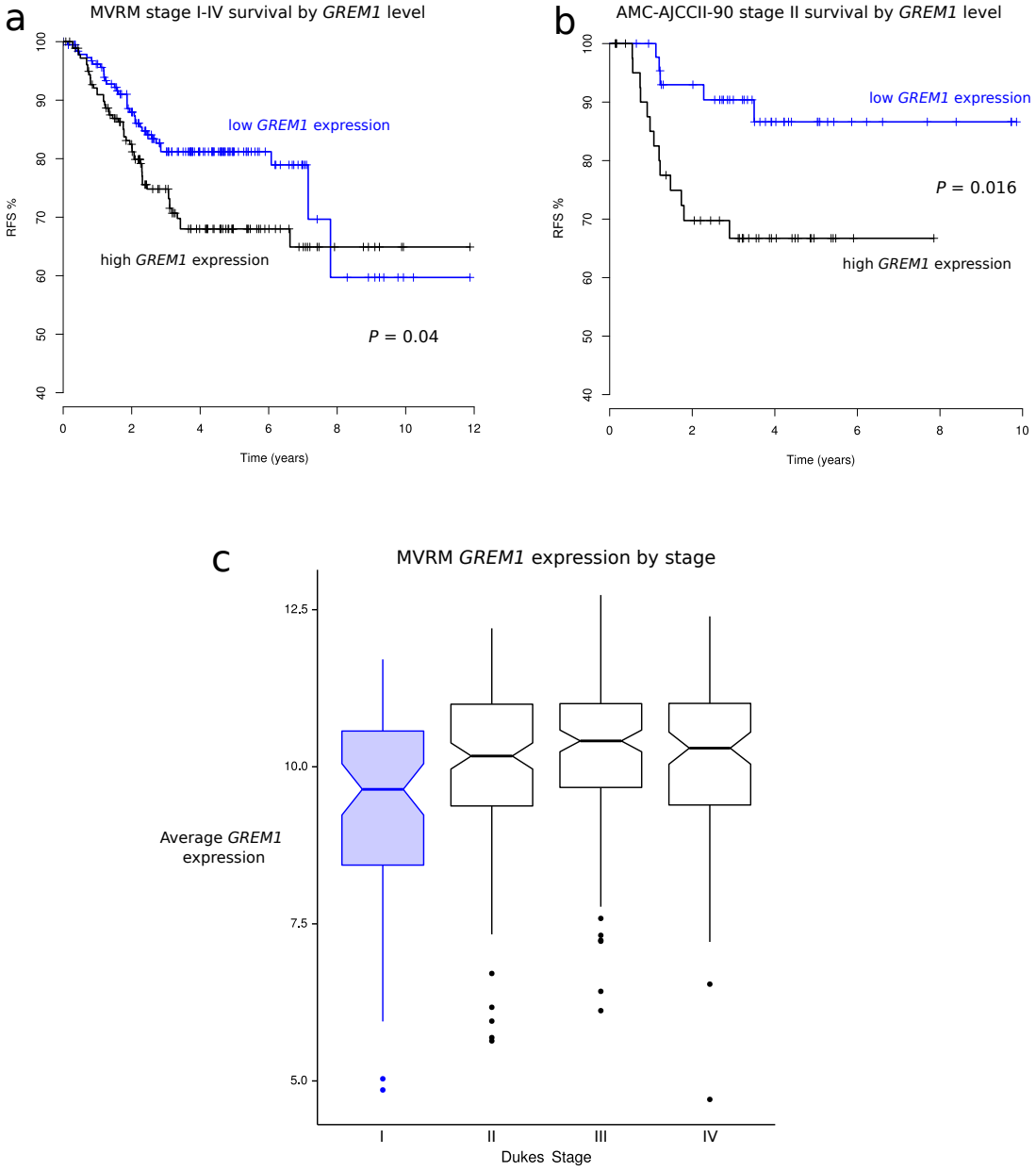


Figure 4.6: *GREM1* expression correlates with survival and stage of CRC. Recurrence-free survival in *a*, the Moffit-Vanderbilt-Royal Melbourne cohort, and *b*, the AMC-AJCCII-90 patient set, stratified by *GREM1* expression (above or below median); log-rank P values are shown. *c*, *GREM1* expression by stage in the MVRM cohort. Stage I expression, blue, is significantly different to every other stage by Tukey's HSD (largest $P = 0.0017$ between stages I and IV). Notches represent 95% C.I.s around the median value.

histograms in figures 4.7b and 4.8b; Shapiro-Wilcox test for deviation from normality $P = 0.26$). It was significantly higher in CRC cases than in controls (figures 4.7b and 4.8b).

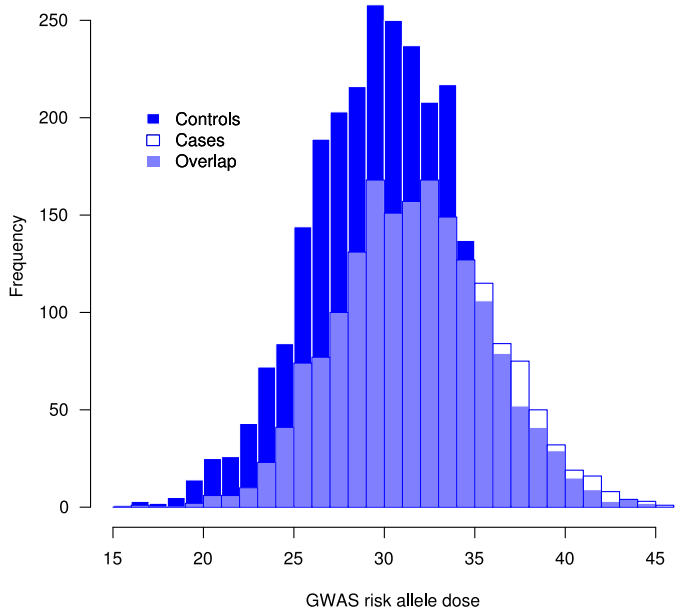
However, unlike *GREM1* expression, the CRC GWAS risk score did not differentiate survival time in either VICTOR and QUASAR2 patients (figure 4.9) or in the TCGA cohort (figure 4.10). Although the lower-scoring median in each group had marginally improved survival, the difference by log-rank test was not significant in any cohort.

4.2.4.1 SNV burden

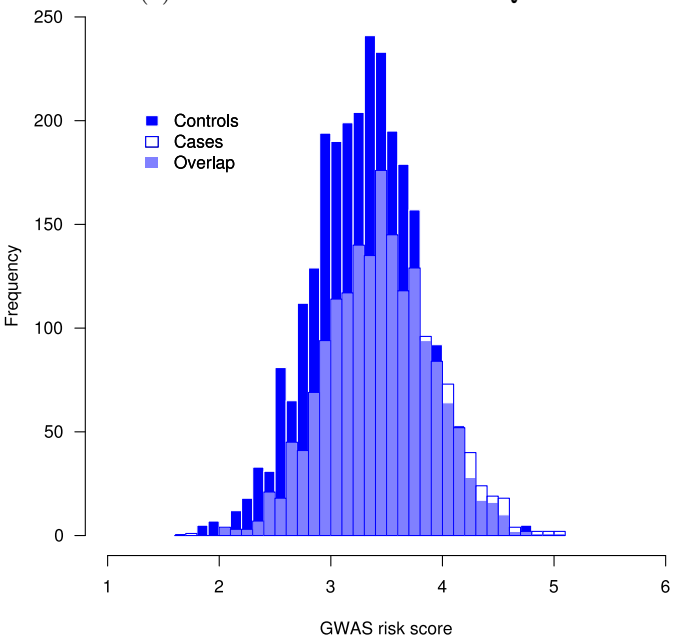
I hypothesised that individuals with a higher germline predisposition to CRC might not require the same degree of tumour driver mutation to develop full disease. To assess whether this was the case, I used two unbiased algorithms that estimate driver status from functional impact bias (Oncodrive-FM, Gonzalez-Perez and Lopez-Bigas, 2012) and positional clustering (Oncodrive-CLUST Tamborero et al., 2013), respectively. These tools are combined in the IntOGen suite (www.intogen.org; Gonzalez-Perez et al., 2013).

I loaded TCGA colon and rectal adenoma sets into IntOGen to generate two lists of driver genes: those with multiple testing-corrected Q values for driver status score was below 0.05 for both algorithms, and those significant in either one or the other. The intersecting, "strict" driver list was composed of 11 genes (*FBXW7*, *TP53*, *APC*, *KRAS*, *CYHR1*, *PIK3CA*, *BRAF*, *TES*, *FAM26F*, *NRAS*, *ZNF439*). The union, "loose" driver list was composed of 458 genes. I then took counts of high-impact somatic mutations within TCGA samples in each list from mutation annotation format files of 270 colon and 116 rectal tumours. These lists were deliberately agnostic to gene familiarity; their purpose was to filter noise from overall mutation burden.

The frequency of driver SNVs within TCGA CRC was right-skewed for both definitions of "driver" (figure 4.11). I therefore plotted a $log(1 + \text{mutant drivers})$ transform-normalised driver SNV count against GWAS risk score for each TCGA CRC sample with both somatic mutation allele frequency and germline genotypes available ($n = 349$), producing figure 4.12. GWAS risk score significantly anticorrelates with number of mutant driver genes, although the slope of this relationship is small ($P = 0.0145$).

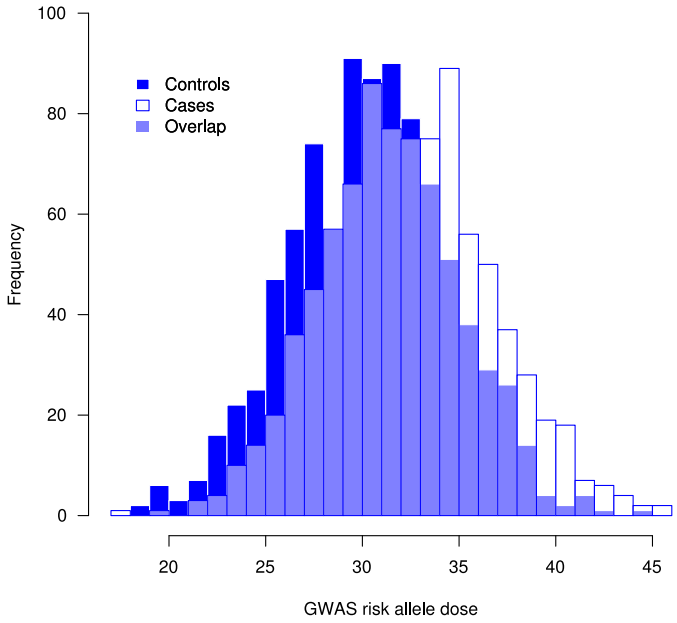


(a) GWAS risk allele dose in VQ58

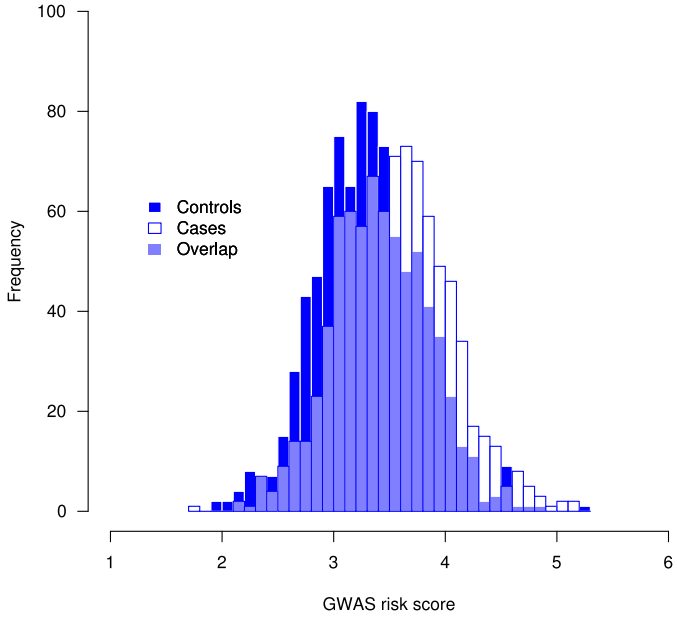


(b) GWAS risk score in VQ58

Figure 4.7: An increased GWAS risk allele dosage and score in VQ58 cases ($n = 1798$) versus controls ($n = 2674$). The differences in means are small but significant ($P_{\text{dose}} = 2.97 \times 10^{-24}$; $P_{\text{score}} = 1.01 \times 10^{-22}$, t tests).



(a) GWAS risk allele dose in CORGI



(b) GWAS risk score in CORGI

Figure 4.8: Histograms of CORGI GWAS risk allele dose and score, showing an increase in both in cases ($n = 888$) compared to controls ($n = 899$). The differences in mean dose and score are significant ($P_{\text{dose}} = 5.94 \times 10^{-23}$; $P_{\text{score}} = 2.26 \times 10^{-21}$, t tests).

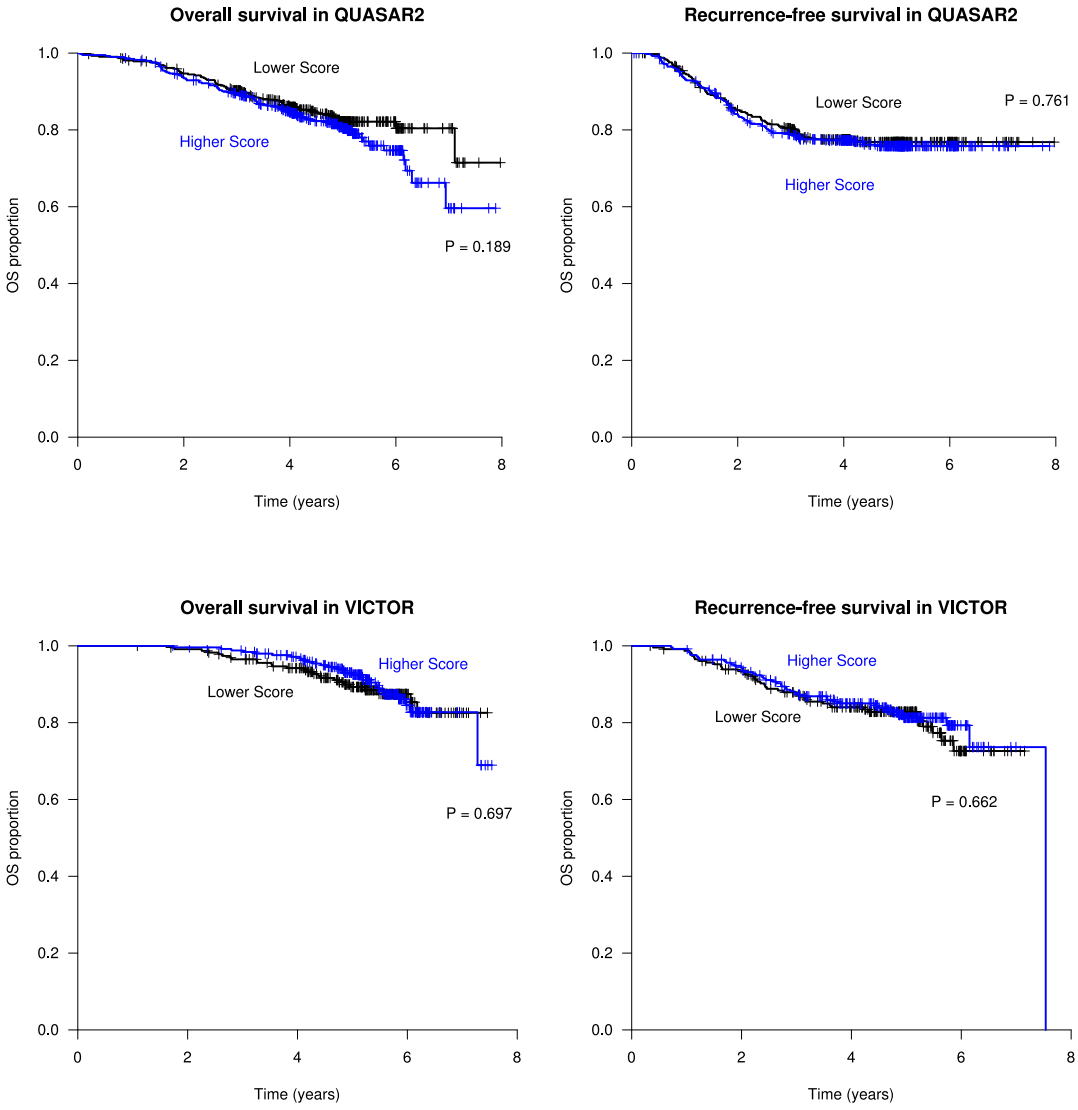


Figure 4.9: Survival curves stratified by risk score for the two cohorts, showing no significant differences (log-rank P value is shown).

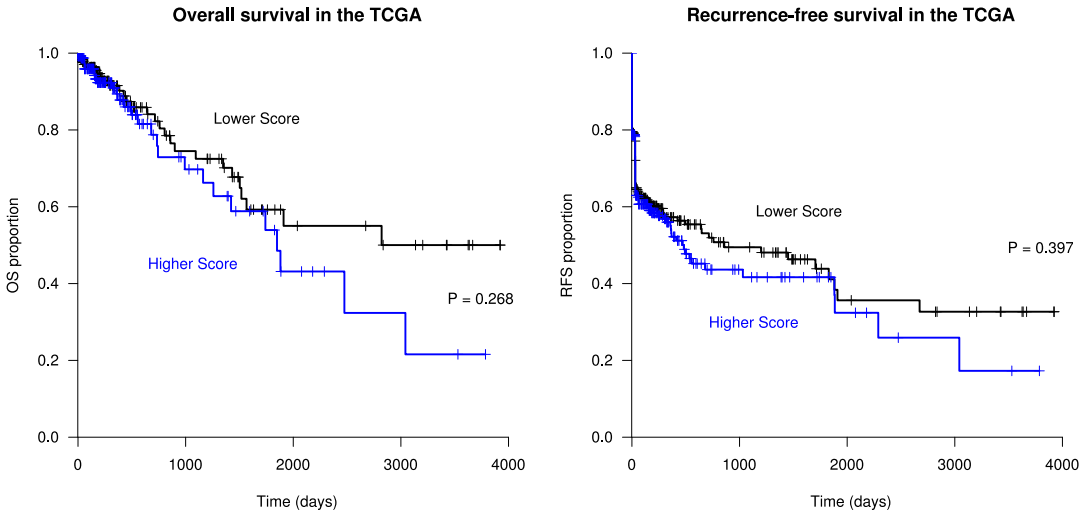


Figure 4.10: Kaplan-Meier curves for overall and recurrence-free survival in TCGA patients, split by GWAS risk score. Log-rank P value is shown.

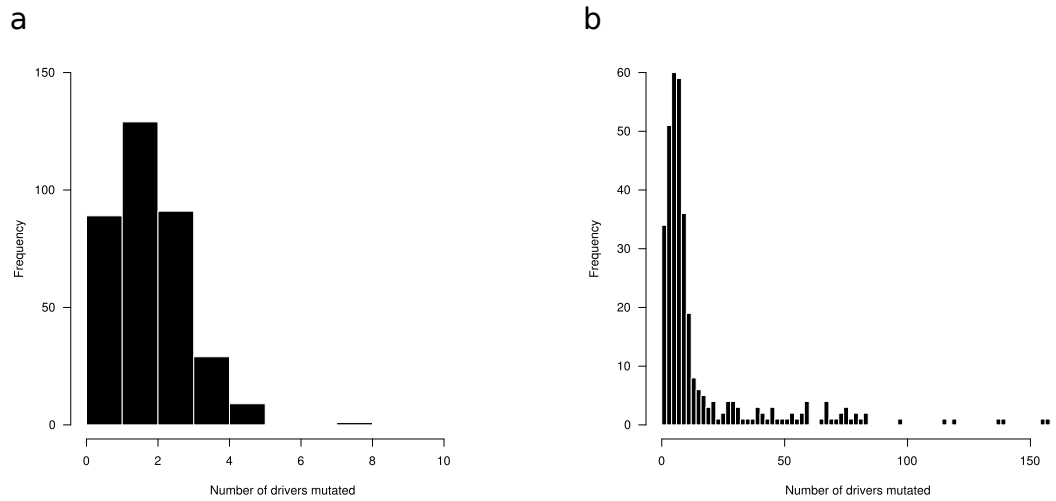


Figure 4.11: Distribution of mutant drivers in TCGA colon and rectal samples. *a*, strict definition (11 drivers); *b*, looser definition (458 drivers)

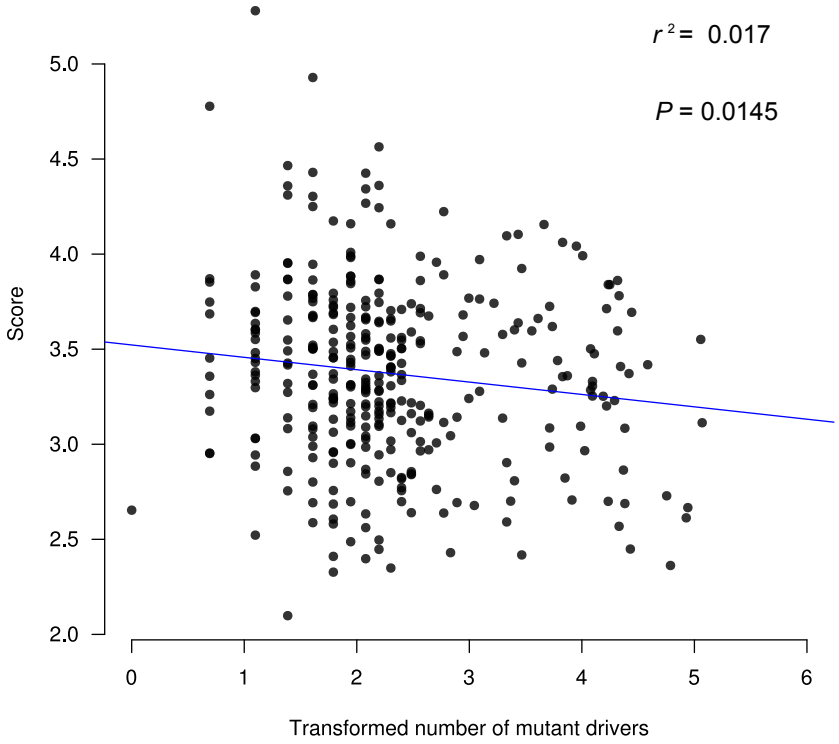


Figure 4.12: Significant but slight correlation between CRC GWAS risk score and $log(1 +$ mutant drivers per individual), with a looser definition of driver gene (see text). This correlation was also significant by nonparametric test ($P_{\text{Spearman's}} = 0.0216$).

4.3 Discussion

GWAS is difficult. Chapter 3 showed the problems of split signals and of the bottleneck of low-throughput functional molecular biology. The exploratory computational work of this chapter reveals at least four additional challenges.

One, LD limits the resolution of fine-mapping, sometimes punitively. Imputation has almost allowed people to forget this, but the ability to improve a signal's P value does not automatically translate to a refinement of that signal. The breadth of candidates at the *GREM1* promoter and (especially) throughout *POLD3* illustrates this. The width as well as the height of local manhattan plots must be taken into account, and haplotypic models more often considered. As I argue in chapter 7, haplotypic risk association is a plausible component of any explanation of "missing heritability".

In particular, the possibility of a haplotypic effect at *POLD3* was previously considered and rejected by Whiffin et al. (2013), despite the haplotypic test's significance ($P = 6.1 \times 10^{-6}$). Not only was the author's own top SNP absent from this test, it also seems that the result was rejected because the P value was larger than the P value of association of rs3824999. But these numbers refer to different statistical questions. Like multi-SNP models, haplotypic effects are more often genuflected at than seriously considered. This is undesirable.

Two, bioinformatic databases are not always helpful. Because many tag SNPs are near genes, and because gene regulatory elements also constellate these spaces, a mere overlay of functional marks with GWAS hits is very likely to provide false-positive explanations. The problem is compounded by the dynamic, tissue-specific nature of many regulatory elements, which explains away a lack of functional correlates at other hits. If a SNP overlaps a DHS or ChIP-seq peak, it could be meaningful; if not, then perhaps its effect is more tissue-specific. It was not considered remotely worrying, for example, that only 5 of 23 reported prostate cancer GWAS hits in Al Olama et al. (2014) were at weak eQTLs in prostate TCGA data, though P values were far from genome-wide significant. This logic is anathema to rigorous hypothesis testing. It becomes impossible to distinguish between false and true negative results when looking for allele-specific

regulatory changes, and is another reason for the genetic evidence to take precedence.

A particular example of this in my work is the absence of a convincing eQTL at 11q13.4. This means that there is still no knockout reason to consider *POLD3* the affected gene. The risk haplotype does span *POLD3*, but could contain regulatory sequences for other regions. The telomeric *CHRDL2* is a perfectly plausible candidate (see figure 4.5). CHRDL2 binds BMPs through its von Willebrand factor type C domain (Zhang et al., 2007), possibly in concert with other BMP antagonists and the modulating TWSG1, to antagonise extracellular BMP signal (Troilo et al., 2015). Public databases of tissue-specific chromatin conformation, or much richer eQTL resources, would both help to substantiate SNP-gene interactions, and perhaps begin to allow the reintroduction of true negative results in genomics.

Three, the utility of GWAS variants as CRC biomarkers is not clear. My individual SNPs predicted CRC survival time in neither the TCGA nor VICTOR-QUASAR2 cohorts. A manual selection of independent inputs to a risk score fared no better. Can GWAS help to stratify CRC patients in the clinic, as it is beginning to in breast cancer?

Four recent predictive models have explored the question. In a large patient set of over 42,000 Scandinavians, Dunlop et al. (2013) genotyped 10 common GWAS SNPs, including an unsplit rs4779584, and *BMP2* and *BMP4* variants. The authors concluded that the discriminatory power of their model remained too weak to be helpful for individual patients. Jung et al. (2015) used 23 SNPs in combination with age, sex, smoking status, fasting glucose, and family history in a Korean cohort. The genetic risk score was imperceptibly superior to traditional risk factor prediction. Similarly, Hsu et al. (2015) constructed a risk score with 27 CRC susceptibility alleles, which provided a very modest (but significant) increase in discriminatory accuracy over family history and risk factors alone. Finally, Ibáñez-Sanz et al. (2017) used 21 SNPs to create an unweighted score, never splitting loci into multiple variants, and adjusting with a propensity score combining age, sex, population structure, and education level. Their combined gene and environmental model produced an area under receiver operating characteristic curve (AUROC, the integral of true positive against false positive classification) of 0.63. However, environmental risk factors and family history produced AUROC=0.61. In all four

examples, therefore, the refinement by GWAS input was marginal at best.

There are protestations that when enough SNPs are added to these sorts of statistics, their predictive power will increase, as it has for breast cancer. One report pushed AUROC to over 0.8 (Wei et al., 2009). However, this was achieved by incorporating interaction terms, for which there is little evidence in either CRC or breast cancer (Ibáñez-Sanz et al., 2017; Kuchenbaecker et al., 2017).

My work in this area had no access to environmental and familial risk factor data and so could not compare traditional and genetic risk scores. This was a major limitation. In my favour were two other differences. First, my SNP score was based on a more thoughtful delineation of independent signals, going beyond the simplistic LD pruning of, for example, Ibáñez-Sanz et al. (2017). Second, I was able to interrogate more than one outcome for association with risk score. I looked not only at case/control status but also at survival and somatic mutation patterns. I could ask whether germline variation at loci associated with CRC formation was also linked to CRC behaviour.

I interrogated two endpoints in particular: survival, and somatic mutation. I uncovered no single SNP association with survival, which was perhaps unsurprising given the absence of *GREM1* or *POLD3* locus survival associations in the larger Morris et al. (2015) study. Nor was my risk score significantly associated with recurrence. As such it appears that CRC GWAS loci are not yet sufficiently numerous to act as clinical prognosticators. However, my score did anticorrelate with somatic driver mutation.

The relationship between germline and somatic CRC genetics has been previously explored: Machiela et al. (2015) looked for an enrichment of somatic mutation in CRC GWAS risk genes, but found none. However, to my knowledge this thesis is the first report of a relationship between GWAS risk score and CRC mutation burden. It closely resembles the recently published work in 638 TCGA breast cancer patients by Zhu et al. (2016). The authors created a weighted polygenic risk score from 90 imputed breast cancer GWAS risk loci, and adjusted for age at diagnosis, stage, oestrogen receptor, and progesterone receptor status. This score had a small but significant anticorrelation with total somatic mutation number. The authors' correction for confounders was an improvement on my work. An extension of my score would incorporate MSS/MSI status

and other covariates, as in Zhu et al. (2016). It would also be interesting to discover whether total mutation counts, as well as my driver subsets, are inversely correlated with risk score. Neverthelss, the similarity of the results is striking.

The sometime influence on survival and driver mutation of GWAS SNPs raises the possibility that the effects of this variation do not disappear upon transformation. This is related to the fourth challenge of GWAS: defining the moment of risk impact during tumourigenesis. Do GWAS SNPs mildly increase the likelihood of adenoma formation, and then disappear? Or do they help tip the scales into full malignancy, or both? By undertaking GWAS for association with adenoma rather than carcinoma, two groups have attempted to unpick these questions. Carvajal-Carmona et al. (2013) found that 8 of 18 tested GWAS SNPs were also significantly associated with adenoma risk. Abulí et al. (2016) genotyped more risk SNPs and saw that 14 of 30 remained significantly associated with the new phenotype. However, the papers are mildly contradictory, with some SNPs significant in one and flat in the other. Five loci are conclusively associated with adenoma as well as carcinoma risk in both: *MYC*, *POU2AF1*, *BMP4*, *SMAD7*, and 10p14 (which lies in a gene desert). The *POLD3* locus was only tested in Abulí et al. (2016) and was declared significantly associated with adenomas by the authors, although its multiple testing-corrected Q value is over a standard threshold of significance ($Q = 0.054$). The *GREM1* SNP rs4779584 was not associated with adenoma risk in either study. Strictly speaking one shouldn't take this as evidence for the null, but this could mean that the *GREM1* risk impinges on colorectal biology after adenoma formation, which is consistent with Lewis et al. (2014) and Pelli et al. (2016). More GWAS of this kind, with more precise phenotypes punctuating CRC progression, would be welcome.

Chapter 5

Population genetics of 461 UK genomes

5.1 Introduction

This thesis will now turn from array and imputed data to whole-genome sequencing (WGS). The Tomlinson laboratory has accumulated a large collection of WGS data on CRC, adenoma, and polyposis patients, some of have whom early-onset or probable syndromic disease (see chapter 6). These samples were collected from British hospitals, and most patients were of recent British ancestry. I unified these genomes under one calling strategy, performed a series of quality control steps, and asked what degree of population structure was retrospectively visible within the cumulative cohort.

The best approach to large sets of WGS data is as much as possible to combine binary sample alignment/map (BAM) files in one joint-calling step. This improves low frequency variant calling and is the only way to be sure of missingness levels at each called variant. Genotype calls are stored in variant call format (VCF). However, traditional joint-calling straight from BAM collections scale poorly in terms of computational cost, and must be repeated every time the sample cohort is expanded. These problems are solved by using a single-sample caller which retains information about reference genotypes, followed by a simpler merge and joint-genotyping step. This was explained in more detail by D'Auwera (2015), who recommended the Broad Institute's software and pipeline, the Genome Analysis Toolkit (GATK). Using GATK tools, single-sample calls can be stored in gVCF rather than VCF files, allowing individual distinctions between missing and reference reads to be preserved. An instant multisample joint-caller (Platypus, Rimmer et al., 2014) and GATK are compared in this thesis.

I discussed in chapter 1 the ability of principal component analysis (PCA) to reveal cryptic population structure (see in particular figure 1.4). This has been used in other large studies to describe recent demographic history. I will discuss three such projects: the Genome of the Netherlands (GoNL), People of the British Isles (PoBI), and the United Kingdom 10,000 genomes (UK10K) project. These three all postdate and complement the 1000 Genomes project (1000G, §1.4.2) as catalogues of variation; the first two were also explicitly interested in using population genetics to examine demography.

GoNL whole-genome sequenced 769 people of Dutch ancestry to a depth of 13× (The Genome of the Netherlands Consortium, 2014). They catalogued some 7.6 million novel SNVs, and created a superior reference panel for rare variants from their phased data. The first principal component of their genotype covariance matrix roughly separated North Holland from South.

PoBI used a 600,000-strong SNP array rather than WGS, and had strict recruitment criteria: UK citizens who were genotyped were required to exhibit a geographically stable recent family history (Leslie et al., 2015). A sampled individual must have lived in the same rural area as all four of their grandparents. Interestingly, despite these criteria, PCA did not reveal much population structure. The first principal component separated out individuals from North Wales, and the second produced a small cluster of people from Orkney or Argyle and Bute. Otherwise, all samples from the other 15 regions of Britain clustered indistinguishably together, regardless of component axis. Leslie et al. (2015) required phasing and hierarchical clustering with the fineSTRUCTURE algorithm to describe finer population structure within their 2039 UK samples, and even then England was one large undifferentiated genetic cluster.

The UK10K project collected nearly 4000 7× read depth whole genomes from two existing longitudinal population studies, and roughly 6000 80× read depth exomes from three disease cohorts. These collections revealed approximately 24 million novel SNVs on top of the 1000 Genomes and GoNL projects (Walter et al., 2015). The task of the UK10K collection was mostly imputation panel creation and association study, and although they did assess population structure, PCA was only used to filter out non-Europeans and was not reported to show any internal structure. The method used in

PoBI, fineSTRUCTURE, revealed no discrete genetic clusters within UK10K. However, very rare variants, with allele counts (AC) between 2 and 7, showed an enrichment cline across geographic distance, with more rare allele sharing between geographically closer samples.

The Tomlinson genomes are a *post hoc* collection of pathogenically selected samples without validated or stable geographic information; they resemble the UK10K collection more than PoBI. In this chapter I will characterise various genomic qualities of the cohort and search for hidden geographic information within the genomes, despite their unpromising provenance.

5.2 Results

5.2.1 Remapping and joint calling

As explained in §2.6.1.1, the 215 Illumina genomes within the Tomlinson cohort required remapping with BWA/Stampy. This increased the mapped read number by an average of approximately 54 million reads ($P = 1.5 \times 10^{-64}$) compared to the Isaac mapping, which represents roughly 4% of the average number of reads in each BAM file.

Two of these remapped case BAM files shared a duplicate identifier and were excluded from further analysis; one was a duplicate sample with a separate identifier and was removed; and two further Illumina samples were duplicates of CGI samples, and were also removed. I then used the GATK pipeline to call 210 Illlumina genomes, 24 "HPPS" samples, and 65 WGS500 CRC cases (the 299 "Tomlinson genomes" hereafter), together with 144 WGS500 controls. These were individuals with extracolonic, non-cancer phenotypes.

The final group of samples were the 203 CGI case genomes, mapped and called with Complete Genomics-exclusive software into the masterVar format, and therefore not amenable to remapping or calling more thoroughly. The "CGI genomes" were used as variant frequency annotators, rather than true cases in their own right, due to this processing constraint.

In parallel, Dr Silvia Salatino multi-sample called the Tomlinson genomes and WGS500

controls directly from BAM files using Platypus. This allowed me to compare the two calling strategies.

Figure 5.1 shows that GATK called more rare and fewer common SNVs and indels. Each sample had fewer non-reference calls after GATK calling than after Platypus calling (table 5.1). However, the combination of these two properties meant that GATK actually called more total unique variants across the Tomlinson cohort than Platypus (32,796,566 versus 30,961,101, respectively). The average sample transition:transversion ratio in the Platypus calls was also remarkably low, at 1.5 (compare to figure 5.2b). The larger number of total and rare variants in the GATK calls encouraged me to continue with this pipeline.

5.2.2 Relative identification by IBD

To properly describe variant frequencies within these sample populations I required a precise understanding of the relatedness of included individuals. I estimated degree of genetic identity by descent (IBD) between all cases using PLINKv1.9 on a small set of independent (unlinked) SNPs from the Omni1M array, extracted by Platypus (see §2.7.6). This identified three relatives within the CGI set, and ten within the other Tomlinson genomes, whom I filtered out (see table 5.2). In each case, I made sure to exclude the relative with the less florid phenotype according to their medical history, be it fewer or later polyps, or later age of onset of CRC. Many of the observed IBD states were confirmed by paper records, such as the sisters `IT-4.1` and `IT-4s`, who share an uncle, `IT-3`, whom I retained.

5.2.3 Basic metrics

Figure 5.2 shows a number of sanity-check metrics for the 289 unrelated Tomlinson genomes called together. The sequencing is of a markedly higher quality than other, larger cohorts such as the UK10K, with an increased average read depth (figure 5.2a) and a reassuring transition:transversion ratio (figure 5.2b) and indel count (figure 5.2d). The GoNL project had an average read depth of 12×, and the UK10K genomes were sequenced at 7×. Presumably as a result of this greater read depth, Tomlinson sample

Type	GATK sample average	Platypus sample average	P value
Reference hom	24,487,934	20,919,375	2.7×10^{-257}
Variant hom	1,449,448	1,533,259	7.8×10^{-169}
Heterozygote	2,384,127	3,998,465	8.0×10^{-153}
Indel	884,030.8	1,294,229.2	3.2×10^{-304}
Singelton	29,104.78	29,268.57	0.95

Table 5.1: Comparison between platypus and GATK multisample calling for 299 Tomlinson genomes. P value represents t test of true difference in means.

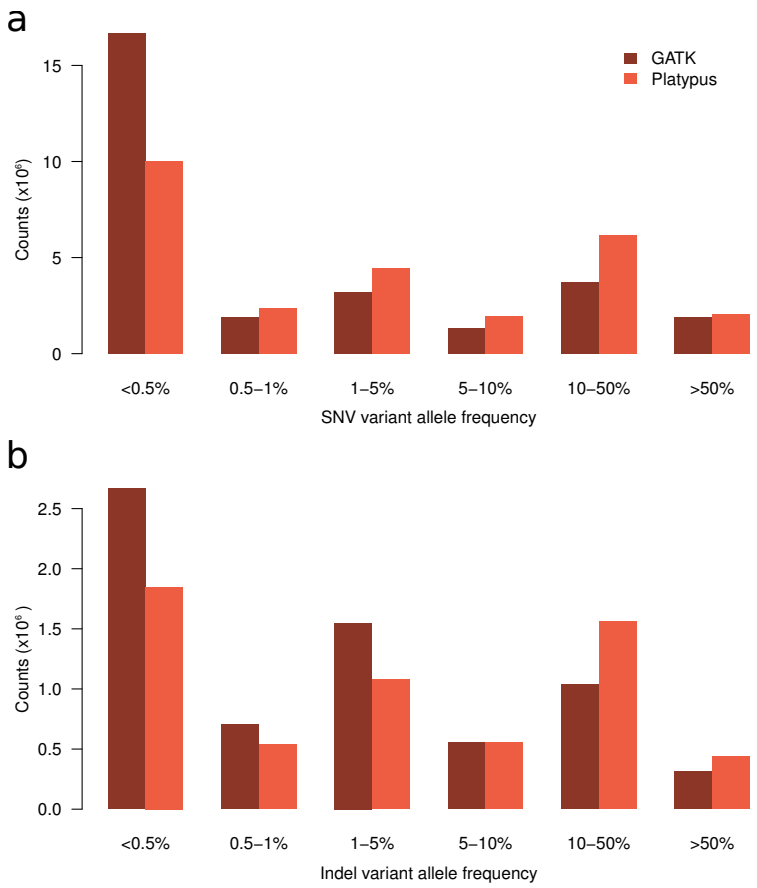


Figure 5.1: A comparison of variant frequencies in the 299 Tomlinson genomes when multi-sample called either with GATK or Platypus software. *a*: GATK calls more rare single nucleotide variants, and fewer frequent SNVs than Platypus; *b*: a similar pattern occurs for indels.

Sample ID	Relationship
GS000004710-ASM	Duplicate sample
GS00416-DNA_G06_1110_37-ASM	1st degree with GS00416-DNA_C05_1110_37-ASM
GS000003862-ASM	3rd degree with GS00416-DNA_H11
LP6005223-DNA_A01	1st degree with LP6005223-DNA_C01
LP6005223-DNA_B01	1st degree with LP6005223-DNA_D01
LP6005223-DNA_H03	2nd degree with LP6005223-DNA_D02
LP6005794-DNA_A01	1st degree with GS000003866-ASM
IT-13	1st degree with IT12
IT-14	1st degree to GS000004212-ASM
IT-4.1	1st degree with IT-4s, 2nd with IT-3, and 3rd with IT38
IT-4s	1st degree with IT-4.1, 2nd with IT-3, and 3rd with IT38
IT_CRC_7	3rd degree with IT_CRC_8
IT-2	3rd degree with IT-1

Table 5.2: People not to include in frequency calculations or PCA eigenvector calculation due to relatedness.

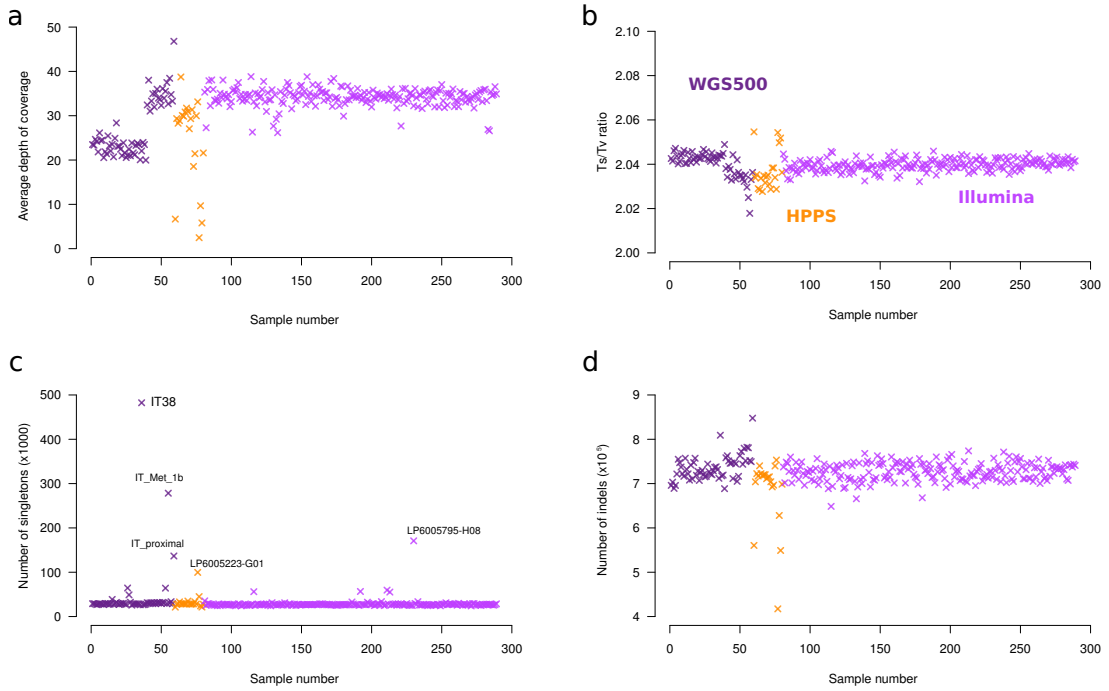


Figure 5.2: Basic metrics of 289 unrelated Tomlinson genomes. *a*: average depth of coverage, which depends on the sequencing cohorts (colours defined in *b*). Most WGS500 samples were sequenced at 25×, and most Illumina at 35×. *b*: a consistent transition/transversion ratio. *c*: disparate degrees of singletons, most helpfully explained by population structure (see §5.2.6). *d*: indel number was steady between samples at around 750,000.

singleton count had a high baseline (at around 30,000 per sample) compared to either of these collections. What stands out in figure 5.2c, however, is the handful of samples with an order of magnitude more singelton variants, not least `IT38`, whose genome contains 482,383 private variants.

5.2.4 Removing individuals with known polyposis mutations

In theory, no Tomlinson sample should suffer a known polyposis syndrome besides PPAP. In practice, the screening for inclusion was imperfect and some genomes may be from individuals with missed polyposis diagnoses. After filtering on allele missingness ($< 25\%$) and heterozygote alternative read depth (> 3), I searched all canonical familial CRC genes for loss of function variants. The resulting list of variants underwent a number of filters before samples harbouring those variants were removed from further analysis. These were: validation by Sanger sequencing, no presence in the healthy database of Bodian et al. (2014), ExAC allele frequency of less than 0.1%, and not an International Society for Gastrointestinal Hereditary Tumours (InSiGHT, Thompson et al., 2014) "variant of unknown significance". Additionally, all polymerase exonuclease domain mutations at positions with functional information, all *APC* frameshift mutations before the gene's mutational cluster region, all InSiGHT class 5 variants, and all splice donor and acceptor mutants were excluded. The individuals thereby removed from the Tomlinson set are shown in table 5.3.

The missed Lynch, FAP, and JPS mutations could reflect the imperfect genotyping techniques of molecular pathology laboratories, which may look for specific, more common mutations. But they are more likely due to conservative eligibility criteria for genetic testing than to pure scientific error.

5.2.5 Other Mendelian mutations within the Tomlinson genomes

I wondered whether the Tomlinson genomes harboured other known Mendelian mutations. This search was interesting for three reasons. First, it might have uncovered genuine, undiagnosed disease. However, this was unlikely due to the recruitment criteria of the sequenced cohorts. Second, it described variation in my genomes in biologically

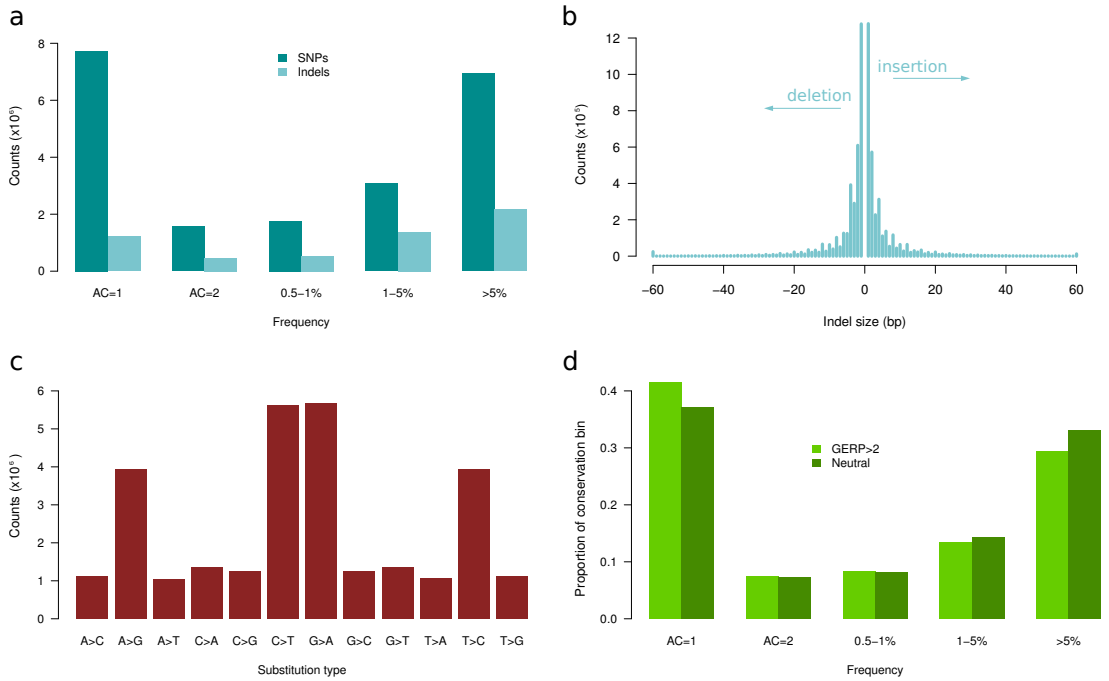


Figure 5.3: More detail on SNVs and indels found within 289 unrelated Tomlinson genomes. *a*, variants binned by derived non-reference allele frequency, including singletons (allele count, AC = 1), doubletons, 0.5–1%, 1–5%, and over 5%. *b*, indel size distribution, where negative lengths represent deletions and positive lengths represent insertions. *c*, base pair substitution type distribution. *d*, proportion of SNPs with either GERP$>$ 2 ("conserved") or GERP$\leq$ 2 ("neutral") split by derived non-reference allele frequency.

chr	Gene	Variant	Effect	ID	Notes
2	*MSH2*	47641560:A>T	splice variant	GS00416-DNA_E01	
12	*POLE*	133250250:G>C	L424V	GS00416-DNA_E05	
18	*SMAD4*	48603147:G>A	splice donor variant	GS00405-DNA_F06	
19	*POLD1*	50909713:G>A	S478N	GS00412-DNA_G01	Relative of IT-14
1	*MUTYH*	45797228:C>T	G382A	LP6005912-H02 LP6005795-E02	Homozygous
		45798475:T>C	Y165C	LP6005795-E02	
2	*MSH6*	48026704:GA>G	frameshift	LP6005795-A01	
		48028225:C>T	R1035*	LP6005794-C09 IT-15	
		48033744:AAAGC>A	frameshift	LP6005795-F08	
	MSH2	47703538:C>T	R680*	LP6005912-C03	
5	*APC*	112102884:A>G	splice acceptor variant	IT-17 LP6005795-C11 LP6005794-A02	
		112163705:T>C	splice donor variant	IT-39	
		112173339:CTT>C	frameshift	IT-37	
12	*POLE*	133250250:G>C	L424V	IT-3 IT-4s IT-4	Related to 4 and 4s Related to 3 and 4 Related to 3 and 4s
		133252321:CAGTCAAAAA>C	inframe deletion	LP6005223-B01	
19	*POLD1*	50909713:G>A	S478N	IT-14 IT-1 IT-2	Relative of GS00412-DNA_G01 Relative of IT-2 Relative of IT-1

Table 5.3: Excluded cases, with mutations in known colorectal polyposis syndrome genes confirmed by Sanger sequencing; CGI (top), other genomes (bottom).

penetrant, loss of function-intolerant genes, and gave a sense of the frequency of this variation. And third, it provided anecdotal evidence of variably penetrant variation even within a tightly curated gene set.

This set was derived from Chen et al. (2016), who published a list of 874 genes implicated in 584 moderate to severe Mendelian diseases. In their set of 589,306 genomes, Chen et al. (2016) discovered thirteen mutant adults who did not display frank pathology.

I searched the 275 unrelated, undiagnosed Tomlinson genomes, which were annotated with WGS500 control and 196 CGI allele frequencies, for variants strongly predicted to cause loss of function in the Chen et al. set. For additional confidence at individual variants, I considered only variant effect on the canonical or best-supported transcript at each gene, and began by filtering on dbSNP clinical annotations, looking for "pathogenic" or "likely pathogenic" variants, and ignoring synonymous variation.

5.2.5.1 Presumed autosomal dominant variants

A number of genes associated with autosomal dominant (AD) disease harboured pathogenic variants in the Tomlinson cohort. See table 5.4. This table does not include heterozygous variants I manually removed because they were only reported to cause recessive disease at their pleiotropic gene: rs104893879 at *WFS1* (causing AR Wolfram syndrome; Inoue et al., 1998); rs104894619 at *PMP22* (Charcot-Marie-Tooth disease type 1A, which is AR; Roa et al., 1993); and rs59885338 at *LMNA* (AR Charcot-Marie-Tooth disease type 2B1; De Sandre-Giovannoli et al., 2002). Additionally, two Tomlinson individuals were heterozygous for a *PITX2* variant (rs6533526) which has been excluded from strict Mendelian pathogenicity by Medina-Trillo et al. (2016) in their exploration of congenital glaucoma. These two individuals were not the same as those with mutant *FOXC1*, precluding compound heterozygosity.

The fact that I was filtering on ClinVar annotation meant that new variants were unlikely to be discovered, and this strategy could not be repeated in chapter 6 to uncover novel pathogenicity. As such I also used the beta VEP plugin LOFTEE (loss of funtion transcript effect estimator, inspired by MacArthur et al., 2012), extracting high confidence loss of function variants in the Chen et al. (2016) gene set. This selected just

Gene	Variant	Associated AD disorder	Reference
BEST1	rs199529046	Best vitelliform macular dystrophy	Unreported variant
CRYAB	rs141638421	dilated cardiomyopathy-1II	Inagaki et al. (2006)
MYH7	rs45496496	dilated cardiomyopathy	Hershberger et al. (2008)
	rs145532615	dilated cardiomyopathy	Unreported variant
KBTBD13	15:65369895:C>T	nemaline myopathy-6	Unreported variant
TGFBR1	9:101908835:A>T	Loeys-Dietz syndrome-1	Unreported variant
TNFRSF13B	rs104894649	common variable immunodeficiency-2	Castigli et al. (2005)
	rs34557412	common variable immunodeficiency-2	Castigli et al. (2005)
	rs72553883	common variable immunodeficiency-2	Castigli et al. (2005)
FOXC1	rs35717904	congenital glaucoma	Medina-Trillo et al. (2016)

Table 5.4: Variants in the Tomlinson genomes likely to cause AD disease. Top, variants present only once. Bottom, variants present more than once: rs104894649 and rs34557412 AC=2, rs72553883 AC=5, and rs35717904 AC=6.

one variant, the *KBTBD13* SNV rs760102500 in `LP6005794-DNA_E02`. This is not the same *KBTBD13* variant as was caught by ClinVar filtering in table 5.4.

5.2.5.2 Presumed autosomal recessive variants

Recessive disease carriers were present in the Tomlinson cohort as one would expect for a European sample group. The sickle cell trait was absent, as was the most common single cause of familial hypercholesterolaemia in Europeans, the *APOB* variant rs5742904. This SNV is rare enough that its absence is unsurprising: the rs5742904 variant allele is present in 0.13% of NHLBI GO Exome Sequencing Project samples (ESP6500). The most common cystic fibrosis *CFTR* mutation, ΔF508 (rs113993960 or rs199826652, the variants are indistinguishable) was present at a frequency of 1 in 50, consistent with previous measurements (de Vries et al., 1997; Estivill et al., 1997). Seven Tomlinson genomes were heterozygous for Factor V Leiden (rs6025; Bertina et al., 1994). Four carried *ATP7B* variants that cause Wilson disease (Cox et al., 2005; Loudianos et al., 1999).

Two classically reported loci were not in the Chen et al. (2016) set, so I searched for variants there manually. These were *APOE*, and the haemochromatosis gene, known as *HFE* or *HLAH*. To genotype the *APOE* alleles $\varepsilon 2$ and $\varepsilon 3$ required phased data and was therefore beyond my capacity. The $\varepsilon 4$ allele, however, is defined by rs429358 genotype and was present in the Tomlinson genomes at a frequency of 11.1%. This allele is associated with Alzheimer's disease risk (Corder et al., 1993); the four Apo$\varepsilon 4$

homozygotes in the Tomlinson set have a roughly 20× increased risk of the disease (Strittmatter, 2012).

The haemochromatosis mutation, rs1800562, was present 43 times in the Tomlinson genomes. This allele frequency is slightly higher than previous reports for Europeans (Lucotte, 2001; Merryweather-Clarke et al., 1997), but is consistent with the possible Celtic origin of the variant allele (Ryan et al., 1998). The Tomlinson cohort included two rs1800562-A homozygotes, who will carry a strong susceptibility to haemochromatosis (Feder et al., 1996), as well as to subtler effects on iron metabolism (Benyamin et al., 2009).

The homozygous Tomlinson samples at *HFE* and *APOE* led me to search for other individuals homozygous for presumed recessive disease-causing alleles. I used the same filters as in §5.2.5.1. The results are shown in table 5.5. None of these variants were caught by filtering on predictive rather than database annotations. Instead LOFTEE isolated just four variants, in two genes (*DOLK* and *GJB2*), with no homozygous mutants for either.

5.2.5.3 Large structural variants

I also performed a scan for larger structural variants in the set. By modifying a script of Dr. Sam Knight, I converted read files into simulated OmniExpress array genotypes, which could be scanned for B-allele frequency and loss of heterozygosity using Nexus. This confirmed the presence in one Tomlinson individual of a heterozygous deletion of the short arm of chromosome 18. See figure 5.4. No other large structural changes were apparent.

5.2.6 Regional variation

Despite our sub-optimal recruitment criteria for the purpose, I next wondered whether there was population structure within the genetic variation in the Tomlinson genomes, as is seen in many of the larger cohorts discussed in §5.1. The simplest way to look for this was with PCA.

I extracted SNPs from the genomes that were present on the Omni1M array, which

Gene	chr	variant	Syndrome	AC	nHoms	Reported?
DPYD	1	rs3918290	DPD	7	1	van Gennip et al. (1994)
DCLRE1C	10	rs41297018	SCID	10	1	Lagresle-Peyrou et al. (2008)
NPHS1	19	36343206:GTCTCTCTC>GTCTCTC	congenital nephrosis	19	1	-
MC1R	16	rs1805007	Skin/hair/eye pigmentation 2	52	2	Sulem et al. (2007)
		rs1805008	Skin/hair/eye pigmentation 2	52	2	Smith et al. (1998)
RAPSN	11	rs45547231	congenital myasthenic syndrome-11	77	7	Müller et al. (2006)
FMO3	1	rs2266780	trimethylaminuria	98	7	Zschocke et al. (1999)
TYR	11	rs1799989	oculocutaneous albinism-IA	60	7	Oetting and King (1993)*
		rs1126809	oculocutaneous albinism-IA	175	26	Hutton and Spritz (2008)

Table 5.5: Likely AR disease variants homozygous in the Tomlinson genomes. nHoms, number of homozygous samples; DPD, Dihydropyrimidine dehydrogenase deficiency; SCID, severe combined immunodeficiency. * Here the authors conclude that the variant is an ineffective polymorphism.

genotypes variation specific to Europeans (The Genome of the Netherlands Consortium, 2014), and removed areas of known long-range LD, as well as the Major Histocompatibility Complex, by converting the recommended coordinates from Price et al. (2008) from build 36 to build 37 using LiftOver. I then took the same SNPs from the 1092 individuals of the 1000G phase 1 (version 3), combined the cohorts, and constructed principal components with the EIGENSTRAT software, using all individuals but first- and second-degree relatives, who were instead projected onto the calculated axes. As expected, this recreated the standard global triangle of the 1000G samples (figure 5.5), with most Tomlinson genomes fitting into the European corner.

This first PCA allowed me to identify individuals of overtly non-British recent ancestry, as shown in figure 5.6. Most obviously, the Tomlinson individuals included one person of African ancestry (`IT38`), two Asian individuals (`GS000004620-ASM` and `LP6005795-DNA_H06`), and a number of samples with mixed principal components, such as the sisters sat halfway between the European and African clusters (`IT-4.1` and `IT-4s`). These findings are pleasing in the context of figure 5.3, as exactly those individuals with an exceptionally high number of singleton variants are those with African ancestry by PCA.

5.2.7 Tomlinson population structure

To examine whether any British population structure might be visible within the Tomlinson genomes, I first removed the 13 individuals who were obviously of different ethnicity from the first PCAs (either with or without 1000G samples included). My imperfect proxy for ancestry was the address given for consent purposes, which I binned into the level 1 Nomenclature of Territorial Units for Statistics (NUTS) regions: nine areas of England, and one each for Wales, Scotland, and Northern Ireland (see figure 5.8). I also therefore masked any individual with an address given that was not British, regardless of PCA placement: there were 29 such samples, some of which overlapped with the PCA outliers, for a total of 40 exclusions. I considered Guernsey (which provided two samples) part of the UK, assigning it the "South West" NUTS region.

A PCA plot of the non-excluded 461 Tomlinson genomes is shown in figure 5.7a.

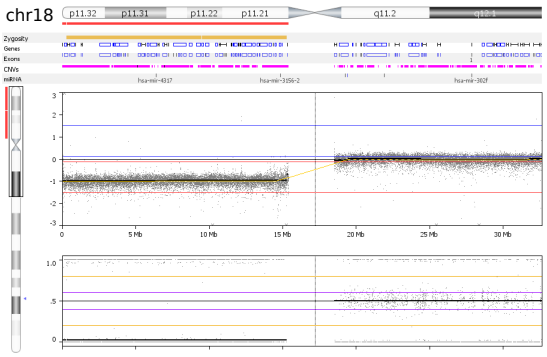


Figure 5.4: One patient lacks an entire short arm of chromosome 18, as seen in Nexus.

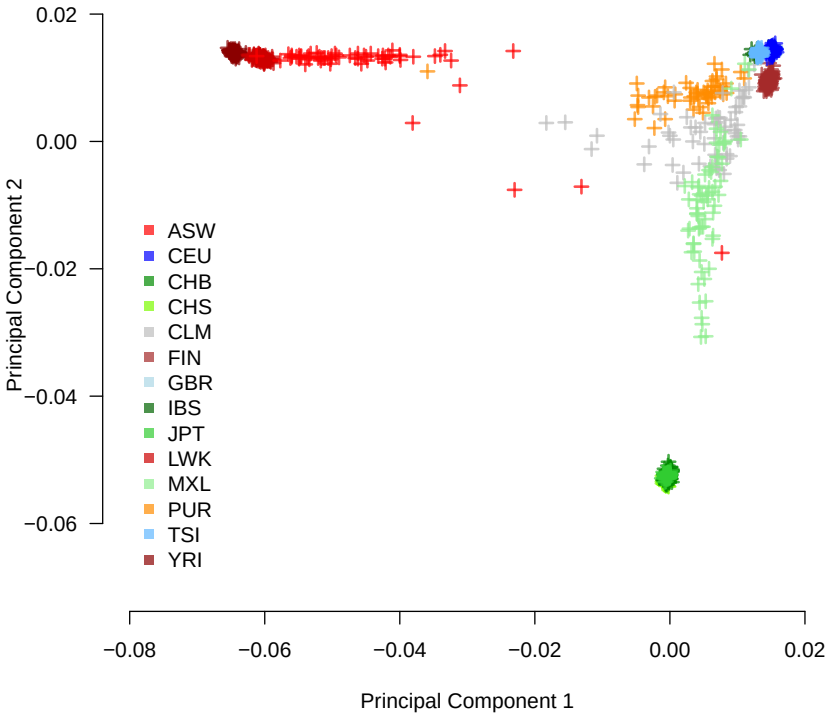


Figure 5.5: Principal component analysis of 1000G samples, coloured by population. ASW, Americans of African ancestry; CEU, Utah Residents with Northern and Western Ancestry; CHB, Han Chinese from Beijing; CHS, Southern Han Chinese; CLM, Colombians from Medellin; FIN, Finnish; GBR, British in England and Scotland; IBS, Spanish Iberians; JPT, Japanese from Tokyo; LWK, Luhya Kenyans from Webuye; MXL, Mexican ancestry from Los Angeles; PUR, Puerto Ricans; TSI, Toscani from Italy; YRI, Yoruba from Ibandan. Note the typical three-cornered structure, with African populations at the top left, Asian populations at the bottom right, and Europe at the hinge.

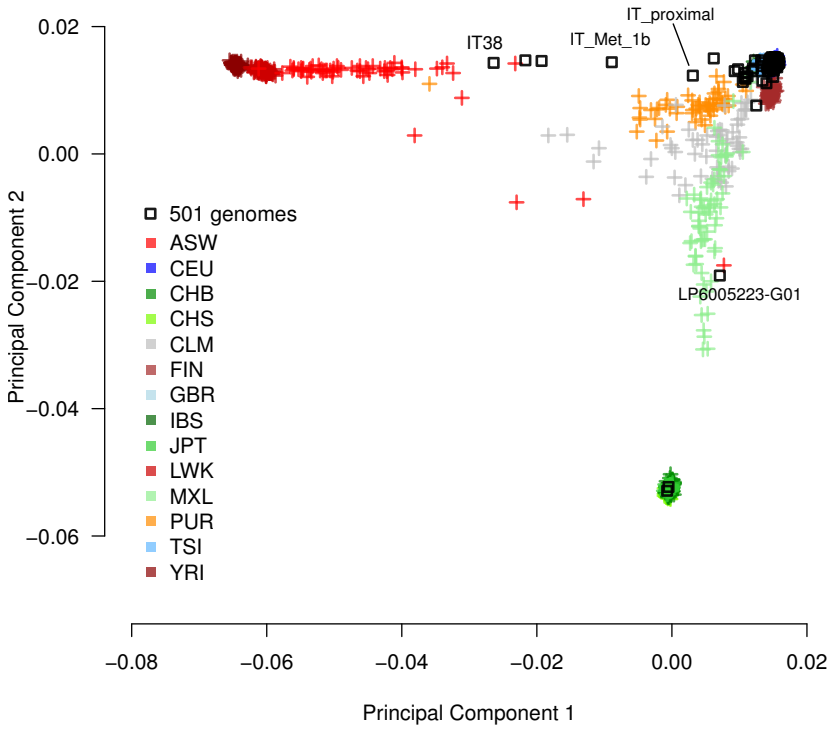


Figure 5.6: 501 Tomlinson genomes combined with 1000G individuals, projected onto the principal components of their unrelated members. Highlighted are some ID codes of Tomlinson genomes of obviously foreign populations.

There is some structure here: most of the genomes cluster tightly, but a small group of individuals from London splits to the left. However, when projected onto pricipal component axes generated exclusively from the 1000G British samples (the "GBR" population), these differences are small (figure 5.7b).

This implies that the Tomlinson genomes are either more homogenous in origin than the 1000G samples, or that they have a simpler information structure. To properly exclude a perceptible regional variation, I needed to look more closely at the dense cluster that contained most of my samples.

I therefore reran EIGENSTRAT with 5 iterations of outlier removal with a σ threshold of 6. This is a complicated way of zooming in on figure 5.7a, leaving 446 genomes. This final plot did at first apear to have a modest West–East polarity, as figure 5.8a shows. However, this may reflect the enrichment of Southwestern samples in the CGI set, which cluster together compared to the other Tomlinson genomes (figure 5.8b).

5.2.8 Rare variants across geography

Obvious population structure was missing in PCA plots of the Tomlinson genomes. I wondered whether, like in the UK10K cohort, rare variants might reflect geography in the absence of coarser PCA clusters.

I therefore took the 238 unrelated, non-CGI samples who were not obviously non-British and for whom some address information was available, and examined their pairwise allele sharing at rare variant loci. Across rare variant frequencies of 2–7 allele counts, there was a consistent excess of rare variant sharing between sample pairs in closer distance bins ($P = 9.15 \times 10^{-7}$; figure 5.9).

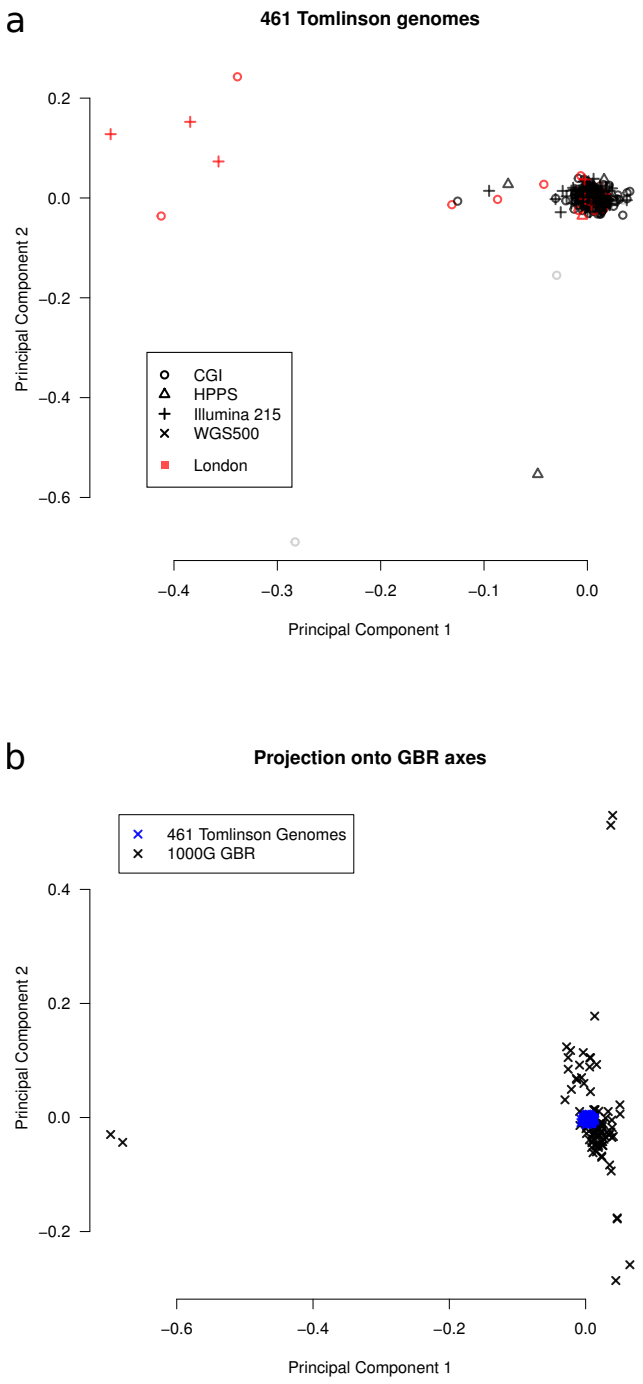


Figure 5.7: *a*, the 461 Tomlinson genomes without obvious non-British members, projected onto their first two eigenvectors. Highlighted in red are a cluster of London samples that do not bunch tightly with the others. The samples to the lower right have no available regional information. *b*, the 461 Tomlinson genomes without obvious non-British members (blue), projected onto the principal components of the 1000G GBR samples (black).

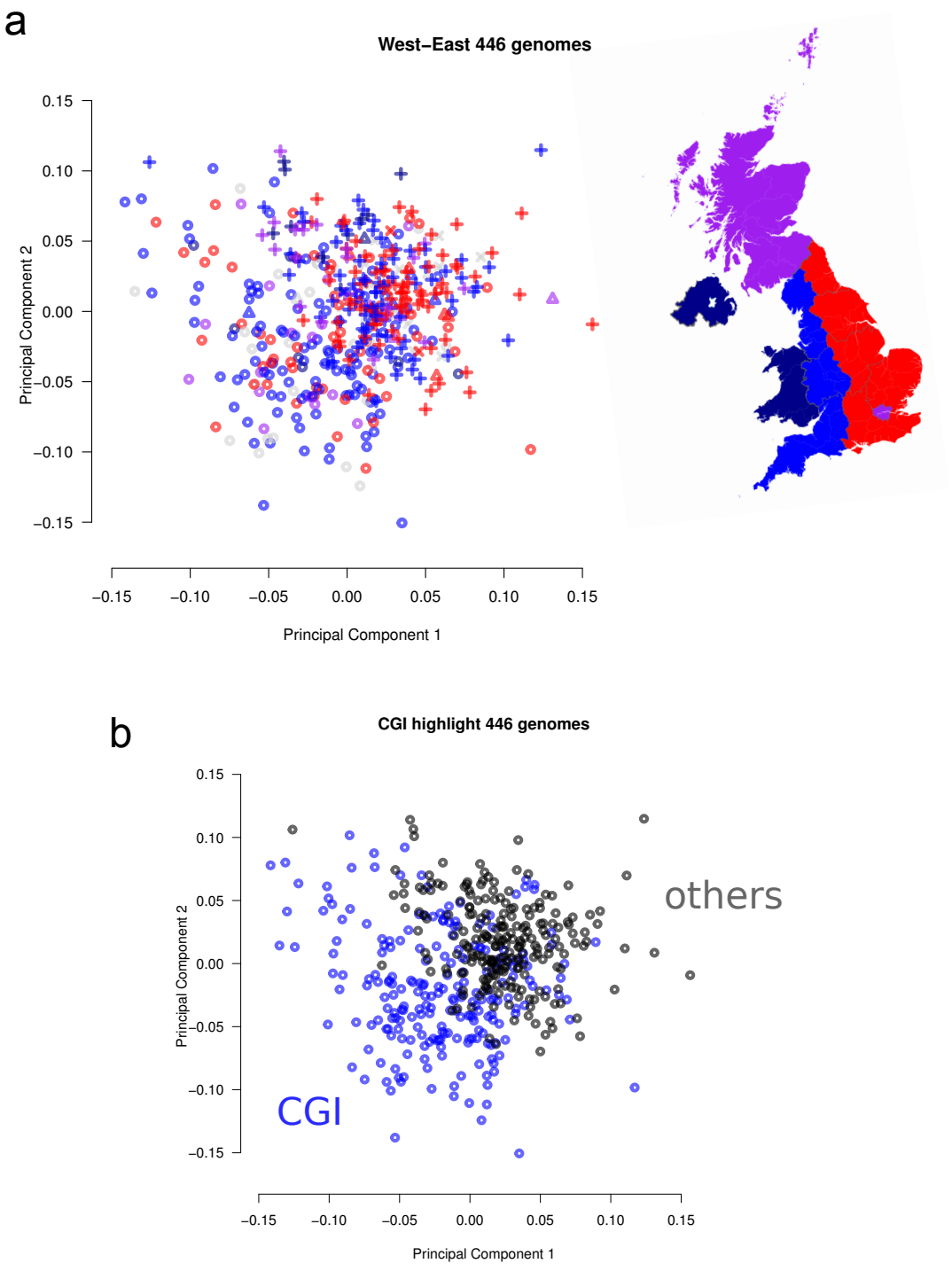


Figure 5.8: Tomlinson genomes PCAs coloured by NUTS region or CGI inclusion. *a*, rough East–West gradient along the cluster of Tomlinson genomes plotted against their first two principal components. NUTS regions of the UK are coloured as shown in map, right. Grey points are of unknown address; point shape corresponds to Tomlinson genome cohort. *b*, a more convincing difference between the CGI genomes and the other cohorts within the Tomlinson set.

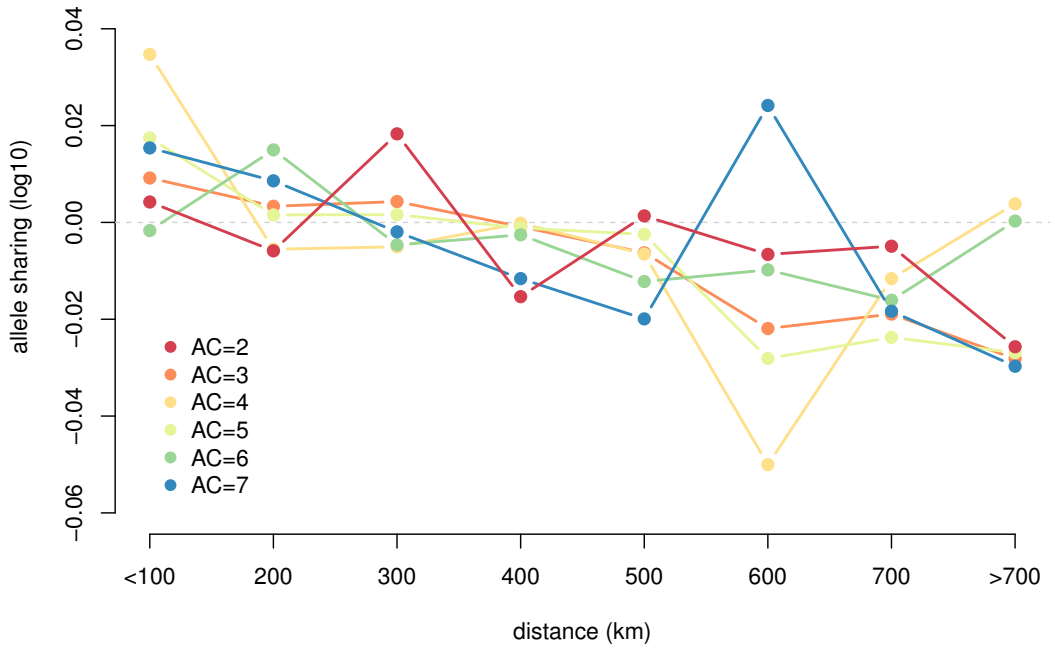


Figure 5.9: Significant excess of allele sharing among 238 non-CGI genomes as a function of geographical distance, expressed as the proportion of shared alleles between sample pairs for total cohort allele count (AC) = 2–7 by distance bin. Overall regression $P = 9.15 \times 10^{-7}$

5.3 Discussion

A whole genome is an overwhelmingly rich source of information; making sense of it can be daunting. As well as introducing and filtering the genomes used in chapter 6, this chapter makes four points of its own.

The first is a methodological argument, which has two parts. One, the GATK multisample calling pipeline is superior to Platypus software. Its calling improvement is slight: in the Tomlinson cohort GATK called marginally more rare variants overall, although per sample it preferentially increased reference allele calls. The real advantage of the GATK approach is that it solves the so-called "$n+1$" problem. Additional genomes can be added quickly to the final multisample callset using GATK; with Platypus the whole calling process must be repeated. Two, downstream annotative tools remain imperfect. The individual variants examined anecdotally in §5.2.5.1 and §5.2.5.2, especially those individually confirmed as pathogenic in the literature, were not uniformly predicted to cause loss of function by VEP. The ΔF508 *CFTR* variant, for example, was predicted to have a "moderate" effect.

Second, the very existence of some of these variants at apparently pathogenic dose is a salutary reminder of variable allelic penetrance, and of the often murky relationship between trait and fitness. The presence of non-colorectal pathogenic variation within the Tomlinson genomes can be explained by placing this variation into three overlapping categories. The first is late-onset phenotype. The Apoε4 homozygotes are virtually guaranteed to develop Alzheimer's dementia by age 80 (Corder et al., 1993), having an odds ratio of disease of about 11 compared to ε3 homozygotes (Bickeböller et al., 1997). But this predisposition may not have yet become clinically apparent. Similarly, the dominant *CRYAB* variant I discovered was reported by Inagaki et al. (2006), who found the same missense change in a 71 year old patient. This individual had developed cardiac symptoms only after her fourth decade. Some of her siblings also had cardiomyopathy, aged 56 and 66; two other siblings had suffered sudden cardiac death aged 60 and 72 (Inagaki et al., 2006). Another example of late-onset disease is the *FOXC1* congenital glaucoma variant. Normally this condition arises in the first decade. But the hypermor-

phic rs35717904 was first described in a patient with unilateral glaucoma, diagnosed at 60 (Medina-Trillo et al., 2016).

The next category is of subclinical, or phenotypically bearable effects. The purest example of this type of "Mendelian" variation in the Tomlinson genomes is the *FMO3* recessive mutation. This variant produces trimethylaminuria (TMA-uria), which also goes by "fish-odour syndrome". People with TMA-uria cannot oxidise the foul-smelling trimethylamine (TMA) into TMA N-oxide, and have a malodorous sweat. However, the degree of TMA oxidation dysfunction, and of smelliness, is variable and sometimes unnoticed. The variant I discovered, rs2266780, more than halves TMA oxidation capacity *in vitro*, but only gives a "transient or mild malodour depending on environmental exposures" (Zschocke et al., 1999). Another example of fully penetrant, but phenotypically mild variation is some of the common variable immune deficiency (CVID)-causing *TNFRSF13B* mutations. In particular, Castigli et al. (2005) reported rs34557412 in three individuals, diagnosed with CVID aged 41, 51, and 17. These people had recurrent sinusitis and upper respiratory tract infections, but did not require immunoglobulin replacement. Or there's the *MC1R* SNPs, rs1805007 and rs1805008. Variant alleles at these loci all but guarantee red hair colour, which is not a phenotype most would consider medically problematic, despite an increased risk of ultraviolet light-associated tumour risk.

However, the *MC1R* variants also belong in the third category, of incompletely penetrant alleles. As well as red hair, variants at rs1805007/8 increase the likelihood of pale skin, freckles, and skin damage from sun exposure (Sulem et al., 2007). But these are semi-Fisherian, complex traits (as, in fact, is hair colour), and the two SNPs are merely instances of a larger collection of inputs. This is also true of the *APOE* SNPs. Their presence is neither necessary nor sufficient for the traits in question. An apposite final example of this category of captured "Mendelian" disease allele is the *DPYD* SNV, rs3918290. This SNV is both incompletely penetrant and pleiotropic. It unreliably causes dihydropyrimidine dehydrogenase deficiency, a heterogeneous disorder involving seizures, mental retardation, and motor abnormalities (Berger et al., 1984). More commonly, however, rs3918290 causes toxic adverse reactions to the 5-fluorouracil class of

chemotherapy drugs. This coincidentally returns us to CRC: the 5-fluorouracil therapy capecitabine is a mainstay of current CRC treatment. rs3918290 and other *DPYD* variants are strongly associated with capecitabine toxicities, ranging from diarrhoea to neutropaenia (Rosmarin et al., 2014, 2015). Unlike the AR dihydropyrimidine dehydrogenase deficiency, 5-fluorouracil toxicity is rs3918290 dose-dependent (Sumi et al., 1998). Neutropaenia is such an undesirable drug reaction that Cortejoso et al. (2016) conclude that it is cost effective to routinely genotype the variant before 5-fluorouracil treatment.

The third argument to make from work presented in this chapter remains related to CRC. I found mutations within known polyposis genes. In the case of PPAP genes *POLE* and *POLD1* this is understandable, since the disorder had not been characterised at the time of recruitment for most individuals (Palles et al., 2013). However, there was still a surprising degree of canonical mutation within the Tomlinson genomes. Hansen et al. (2017) reported a similar finding in 274 familial CRC patients who had previously screened for MMR and other CRC predisposition genes. Their cohort resembled the Tomlinson set, and was genotyped on a custom panel of 112 known and candidate CRC genes. In 6% of these patients lay undiscovered pathogenic mutations within known CRC pathogenic genes. It's not clear exactly which denominator to use for the Tomlinson genomes, but roughly that proportion of our samples also had "known" genetic disease. I argued in chapter 3 that known GWAS association loci might account for more sporadic CRC heritability than we currently assume. Analogously, it's possible that missed mutation at known polyposis and Lynch genes causes an overestimation of remaining unexplained Mendelian or quasi-Mendelian CRC genetics.

Fourth, despite being unplanned, unstable, and sometimes vague, the geography of the Tomlinson genomes did eventually became visible by variant distribution. Moreover, the 238 individuals used to create figure 5.9 were principally from within a homogeneous PoBI region ("England"), yet had variegated rare allele sharing. This agrees with Walter et al. (2015), who found that rare allele sharing mirrored fine-resolution local geography, producing clines much like the Tomlinson genomes. The implications of this are important, and were established in simulation studies by Mathieson and McVean (2012).

Briefly, rare variant association studies will need to account for this local sharing, which may require new statistical tools. In particular, rare variant association will inflate if nongenetic risk is highly localised, even if the variants are binned into gene groups. This nongenetic risk might be due to different recruitment criteria, or might reflect true local environmental differences such as those caused by pollution. PCA does not work well as a corrective against this inflation because the distribution of a local confounder is highly nonlinear; the amount of PC correction required would effect too large a diminishment of discovery power. Hence, rare variants are an excellent way to see high-resolution population structure that we currently lack the statistical tools to correct for. They represent a challenge to the field.

Chapter 6

Searching whole genomes for pathogenic germline variation

6.1 Introduction

The Tomlinson genomes introduced and characterised in chapter 5 are too few and too heterogeneous in phenotype to power rare variant association study. As such, I must follow the example of the literature discussed in §1.5 and attempt to discover pathogenic variation by filtering on frequency, predicted effect, and gene set. This is a retreat into more Mendelian genetics, as I have pointed out; a retreat that I reinforce by returning to frank linkage study, in the manner of Palles et al. (2013) and (to an extent) of Seguí et al. (2015).

One gene set — that of mismatch repair (MMR) genes — is prompted by Lynch syndrome pathogenesis. And one linkage study is in a pedigree suffering from serrated polyposis syndrome (SPS).

6.1.1 Mismatch repair and Lynch syndrome

The first modern description of a hereditary cancer syndrome was made by Aldred Warthin in 1913, following 18 years of history collation at Michigan hospitals. He described related patients who developed cancers of "the uterus, breast, gastrointestinal tract and mouth [...] The incidence of cancer in these families is so striking that it can be interpreted as showing an inherited susceptibility to cancer" (Warthin, 1913). This extended family was followed for many years, latterly by Henry Lynch (Lynch

and Krush, 1971), who became eponymous with the autosomal dominant predisposition syndrome.

As discussed in §1.2, Lynch syndrome causes cancers in multiple organs, especially the colon and endometrium, without prior polyposis. The disease was mapped and molecularly characterised in the 90s: four genes are mutant in Lynch syndrome. These are *MLH1* (Bronner et al., 1994; Lindblom et al., 1993), *MSH2* (Fishel et al., 1993; Peltomäki et al., 1993), *MSH6* (Miyaki et al., 1997), and *PMS2* (Nicolaides et al., 1994). The genes are all involved in DNA MMR.

MMR in eukaryotes is initiated by the binding to mismatch of one of two heterodimer rings: MSH2:MSH6 or MSH2:MSH3. These then recruit another dimer, MLH1:PMS1, and begin the process of mismatch excision and repair. The MSH2:MSH6 dimer clamps single base pair mismatches and indels; the MSH2:MSH3 dimer instead is recruited to double mispairs and to indels over 2bp (Srivatsan et al., 2014).

I mentioned in §1.1.1 that *MLH1* promoter hypermethylation is a common cause of MSI in sporadic cancers. There is a particular kind of MSI called "elevated microsatellite instability at selected tetranucleotide repeats" (EMAST), which is $[AAAG]_{(n)}$ or $[ATAG]_{(n)}$ instability. This size of error falls under the purvue of MSH2:MSH3 repair, and correspondingly Haugen et al. (2008) found that knockdown or knockout of *MSH3* causes EMAST in cell lines. Lower *MSH3* expression, but not its somatic mutation, is correlated with EMAST in CRC biopsy (Haugen et al., 2008; Plaschke et al., 2012). Overexpression of *MSH3* can also cause MSI (Marra et al., 1998).

In 2016, Adam et al. exome sequenced 102 individuals with unexplained adenomatous polyposis. This revealed two unrelated females with compound heterozygous loss of function mutations in *MSH3*, each with an affected sibling with the same genetics. These mutations survived in mRNA, but knocked out the nuclear protein, and caused an EMAST MSI mutation spectrum in their adenomas. Both individuals had extensive extracolonic neoplastic changes, including thyroid adenomas, breast tumours, and uterine growths including fibroids (Adam et al., 2016). Additionally, one patient developed a teratoma, and the other had an astrocytoma, which is not uncommon in Lynch patients.

6.1.2 SPS

In 1998, Vogelstein proposed the adenoma-to-carcinoma sequence: serial mutation acquisitions transform adenomatous colorectal polyps into tumours (Vogelstein et al., 1988, see §1.1). But not all polyps are adenomatous. Rather than being dysplastic — dedifferentiated adenomas — or improperly developed — the mutiple ectopic tissue types of hamartomatous polyps — some polyps can form from simple overgrowth. These are the hyperplastic, "serrated" polyps. They are now recognised to be a family of three types. The "serrated" epithet derives from their characteristic saw-tooth crypt corrugation under microscopy.

The three kinds of serrated polyp are the traditional serrated, the sessile serrated, and the hyperplastic polyp (Snover et al., 2010). Traditional serrated adenomas (TSAs) are the rarest form. They have a left-sided preponderance and exhibit ectopic crypts, which is reminiscent of the budding histopathology of HMPS (§1.2.3). Indeed, the suggestion of Davis et al. (2015) is that TSAs are the sporadic analogue of HMPS. Epithelial *GREM1* overexpression creates ectopic crypts and polyposis in mice, and *in situ* hybridisation and qRT-PCR show epithelial *GREM1* transcription in human TSA samples.

Hyperplastic polyps are more often seen in the left colon and rectum, are sessile or slightly elevated, and have elongated crypts which are serrated towards the top (Aust and Baretton, 2010). They are by far the most common serrated polyp type: roughly three quarters of reported serrated polyps are hyperplastic (Higuchi et al., 2005). The rarer sessile serrated polyps (or adenomas, SSAs) are, by contrast, more likely to be right-sided, and to have warped, branching, or dilated crypts with extensive lower serration (Aust and Baretton, 2010). Microscopically therefore the two more common serrated polyp types are distinguishable, although on endoscopy they are easy to confuse — or to miss altogether, since both lie flat on the colon wall (Payne et al., 2014).

In 1970, Goldman et al. reported a 42 year old male with 30 serrated polyps. The authors claimed that they saw neoplastic change within these polyps. But this was contrary to orthodoxy: hyperplastic polyps, as they were then exclusively known, were "unrelated as a precursor tissue to either adenomas or carcinomas" (Lane, 1976). It has

taken decades for this idea to be overturned, and for SSAs and serrated polyps more widely to be recognised as cancer precursors. The World Health Organisation (WHO) now defines a serrated polyposis syndrome (SPS) on the basis of three alternative criteria (Snover et al., 2010): *i.*) at least five serrated polyps proximal to the sigmoid colon, with two greater than 10mm across; *ii.*) any number of serrated polyps proximal to the sigmoid colon in an individual with a first-degree relative with SPS; or *iii.*) at least 20 serrated polyps of any size throughout the colon.

Taken as an imperfect group, serrated polyps are in fact associated with an increased risk of CRC (Boparai et al., 2010; Hyman et al., 2004; Leggett et al., 2001), though the strength and nature of this association is inconsistent across what remain mostly small prospective studies. The predisposition is shared by both first- and second-degree relatives of those with SPS (Win et al., 2012), and is to a molecular subtype of CRC with poor prognosis (De Sousa E Melo et al., 2013).

The historical reluctance to accept serrated polyps as conferring CRC risk may have come from two places. First, the Vogelstein model of carcinoma formation was dominant. It took the finding by Jass et al. (2000) of a hyperplastic polyp that was also microsatellite unstable to nudge the orthodoxy towards accepting multiple early routes to bowel tumour.

Second, the grouping of all types of serrated polyp — indeed, the very existence of these categories in particular — was, and remains, complicated. I have stated that TSAs might involve a unique, HMPS-like, tumourigenic progression compared to both tubular adenomas and to other serrated polyp types. Whether hyperplastic and sessile serrated polyps are on a continuum or represent different pathologies is also unclear (Bettington et al., 2013). The two are very frequently confused by pathologists, or else the terms are used interchangeably (Glatz et al., 2007; Khalid et al., 2009). Despite this, all three serrated polyps remain grouped together in the current definition of SPS. Therefore, the possibility of hyperplastic polyps being entirely benign might still mask or muddy real risk from SSAs, as Kalady et al. (2011) imply.

Though the genetics and molecular pathway of SPS-associated CRC risk are uncertain, the field now agrees that this risk is real and warrants the close endoscopic

chr	Gene	Variant	Effect
3	*CTNN1B*	41280814:G>A	W776*
10	*ABCC2*	101591826:C>T	R1066*

Table 6.1: High-confidence effective variants in GWAS genes. Both variants are present only once in the combined Tomlinson genomes and exomes.

surveillance of affected patients (Edelstein et al., 2013; Hazewinkel et al., 2014). The genetic cause of SPS is currently unknown, although two studies have claimed a role for *RNF43* (Gala et al., 2014; Taupin et al., 2015).

In this chapter, I begin by searching an expanded set of both the Tomlinson genomes and a collection of exomes for pathogenic variation in CRC GWAS-associated and MMR genes. Returning to the genomes alone, I ask to what degree noncoding variation can be interrogated by searching for HMPS-like duplications, and by using ChIP-seq data generated in a noncancerous colon cell line. I then take a more careful look at phenotypically precise subsets of the Tomlinson cohort, before performing preliminary linkage study on two pedigrees with unexplained syndromic polyposis, including SPS.

6.2 Results

6.2.1 Gene subsets

The first search for pathogenic variation was not in the Tomlinson genomes alone. They were supplemented by a collection of case exome sequences, which are catalogued in §2.5. In total the coding sequences of 513 individuals were examined.

6.2.1.1 GWAS genes

I searched the genes within the 31 loci of table 1.4 for VEP-predicted high confidence effect variants, filtering on heterozygote alternate allele read depth, and excluding splice site variants tagged "low confidence" by LOFTEE. This filter left just two variants, shown in table 6.1.

The β-catenin variant is a stop-gain mutation. It's normally gain rather than loss of function in β-catenin that is pathogenic, as the gene is oncogenic (see review by Basu et al., 2016); this variant is unlikely to cause disease.

chr	Gene	Variant	AC	Effect
1	*MUTYH*	45796868:TG>T	1	frameshift
1	*EXO1*	242052846:G>T	1	E829*
2	*MSH6*	48027656:A>AT	1	frameshift
		48028228:CTG>C	1	frameshift
		48033708:A>AATCTCCCAGAGGAAGTT	1	frameshift
3	*MLH1*	37067132:T>A	1	L348*
		37092123:CAA>C	1	frameshift
4	*RFC1*	39314488:G>A	1	R423*
5	*MSH3*	79968072:C>T	2	R268*
7	*PMS2*	6048650:T>C	1	start lost
14	*MLH3*	75515731:G>A	1	R210*
19	*POLD1*	50917091:G>GGT	1	frameshift
X	*RPA4*	96139753:AG>CT	1	E149*

Table 6.2: High-confidence effective variants in MMR genes.

This leaves the *ABCC2* exon 23 arginine 1066 stop-gain mutation, which is catalogued in dbSNP as rs72558199. One exome displays the variant heterozygously. rs72558199 is rare, with the T allele present in ExAC at a frequency of 0.035%. The variant produces a protein lacking an ATP casette and 4 membrane-spanning domains. rs72558199 has been previously connected to disease, although not to cancer. Pacifico et al. (2010) described two brothers with neonatal-onset Dubin-Johnson syndrome, a rare conjugated hyperbilirubinaemia. The brothers had compound heterozygous *ABCC2* loss of function, with one rs72558199 variant allele and one rs56199535 variant allele. The latter causes R768W missense as first described by Wada et al. (1998). The rs72558199 heterozygote my screen uncovered is therefore more likely to be another example of an AR disease carrier, rather than the sufferer of a new colorectal syndrome.

6.2.1.2 MMR genes

The four canonical Lynch genes are all involved in mismatch repair. MMR is further highlighted by the recent discovery by Adam et al. (2016) of a biallelic *MSH3* mutation, also affecting MMR, in an adenomatous polyposis patient. I therefore searched all Tomlinson genomes and exomes for loss of function variants in the 23 genes of the KEGG MMR ontology, using the same (stringent) criteria for variant effect as in §6.2.1.1. The results are shown in table 6.2.

Some of these alleles had been previously picked up when screening for Lynch mutations as part of chapter 5. The *MSH6* 48033708:A>AATCTCCCAGAGGAAGTT variant had not led to sample exclusion because it was not validated by Sanger sequencing. *PMS2* 6048650:T>C, catalogued as rs587779333, was Sanger validated, but is a start-loss missense change in only some transcripts of the gene. In others it merely modifies the 5' sequence. In the InSiGHT database it is placed into class 3 ("uncertain"), having been reported twice before (microattribution ID 1000399).

The *MUTYH* heterozygote, an exome-sequenced SPS individual, axiomatically cannot suffer the recessive *MUTYH*-associated polyposis (MAP), and their disease is therefore not explained by the frameshift.

More plausibly pathogenic are the *MLH1* variants. The tyrosine 509 37092123:CAA>C frameshift change was Sanger validated, but occurs in the nineteenth and last exon of *MLH1*, is "of uncertain clinical significance" in ClinVar, and is not reported in InSiGHT. For these reasons the individual was not removed from the Tomlinson set. However, on closer inspection the frameshift is very similar to the InSiGHT class 5 rs786204318 (K510*). The frameshift deletion at Y509 introduces a STOP three codons later (YSL*). Therefore 37092123:CAA>C is in fact likely to be disease-causing.

37067132:T>A resembles the ExAC leucine-348 to serine missense rs755401753, but with a different variant allele (T rather than C). rs755401753 has been reported in InSiGHT (microattribution 1000245), although it was not assigned a class. But a stop-gain is destructive to a much greater degree. This is the first report of *MLH1* 37067132:T>A, which is likely to be pathogenic.

The *EXO1*, *POLD1*, and *RPA4* variants are also promising. EXO1 interacts with MLH1 and causes MSI when knocked out (Bardwell et al., 2004). The exon 16 stop gain at glutamine 829 (chr1:242052846:G>T, rs757677420) is extremely rare in ExAC (seen just twice, a frequency of 0.0016%). In *POLD1* I found a Sanger-validated, previously unreported frameshift in exon 19, 50917091:G>GGT. This variant lies in the gene's polymerase domain. Although it is uncertain what impact this has on function, similar polymerase domain variants have been seen in other CRC patients (Dr. Claire Palles, personal communication). In the single-exon, X chromosome gene *RPA4*, I report a

double nucleotide variant (96139753:AG>CT, which shares its first alternate allele with the rare SNV rs748956801). The first base is a synonymous change in leucine 148 of the protein, but the second base is a stop-gain missense in glutamate 149; the overall protein sequence of *RPA4* is 261 amino acids long. These last two mutations in particular are plausibly functional and disease-causing.

Most exciting however was the homozygous *MSH3* stop-gain mutation at chr5:79968072. This occurs in exon 5 of 24 of *MSH3*, turning arginine 268 to STOP.

6.2.1.3 The *MSH3* homozygous mutant

The individual with two *MSH3* loss of function changes is a 76 year old female with a history of CRC and mild polyposis. In 2003, aged 62, she had excised a stage pT2, grade 3 rectal adenocarcinoma. Subsequent endoscopies revealed three tubular adenomas at the hepatic flexure in 2009, and one caecal and one rectal tubular adenoma in 2012. The individual is also suspected of having a ruptured mucinous pancreatic cyst, which was last investigated in 2015. The 2003 tumour was MLH1 positive, SMAD4 positive, β-catenin low, lacking vascular invasion, and diploid.

The homozygous *MSH3* mutation was present by Sanger sequencing in both normal and cancer tissue (figure 6.1). Somewhat unexpectedly, tumour DNA was microsatellite stable at all mononucleotide markers tested (see §2.8.7). However, two of five tetranucleotide markers were clearly unstable in tumour compared to normal DNA, as shown in figure 6.2, making the rectal adenocarcinoma EMAST-positive.

6.2.2 Noncoding changes in Tomlinson genomes

I examined the relationship between CRC and intergenic DNA variation in the Tomlinson genomes in two ways. One, I asked whether any other BMP inhibitor genes might harbour large upstream duplications in the manner of *GREM1* in HMPS. Two, I compiled regions of colorectal BMP-responsive SMAD1/5-binding by ChIP-seq and assessed whether rare germline mutation was enriched at these sites.

I searched all DAN family inhibitors and other genes reported to inhibit BMP activity, as shown in table 6.3, for copy-number change using a modified script by Dr.

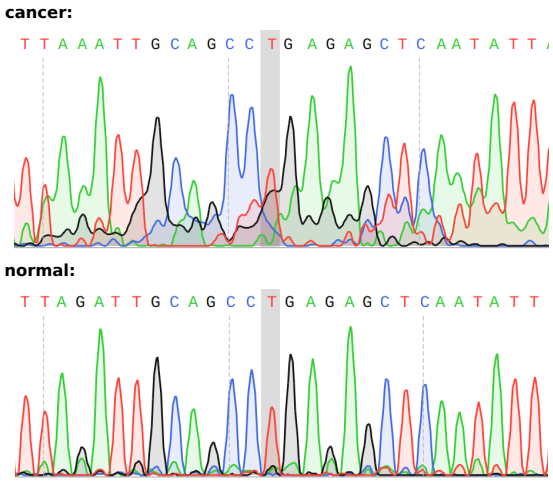


Figure 6.1: The chr5:79968072:C>T *MSH3* mutation, grey stripe, in two sequencing traces from Mseq.Ha47N DNA. Top, tumour; bottom, normal muscularis tissue DNA.

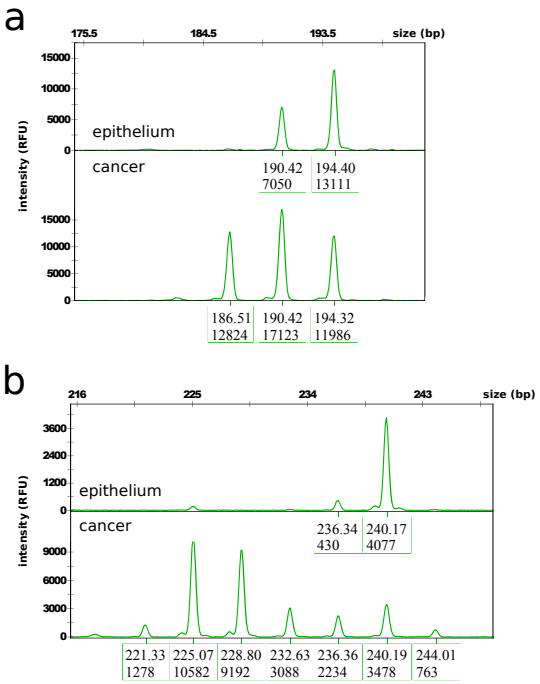


Figure 6.2: EMAST electropherograms of *MSH3*-mutant tumour and normal samples, showing tetranucleotide instability at *a*, D9S747; and *b*, D20S82. Peaks are annotated with their mean size in bp (top number) and their fluorescence intensity (bottom number).

Gene	Name
NBL1	Dan
GREM1	Gremlin1
PRDC	Gremlin2
DAND5	Dante
SOST	Sclerostin
CHRD	Chordin
FST	Follistatin
FSTL3	Follistatin-like 3
TWSG1	Twisted Gastrulation
VWC2	Brorin
BMPER	BMP binding endothelial regulator
DKK1	Dickkopf-related protein 1

Table 6.3: Genes searched for upstream duplications. Top, DAN family. Bottom, other BMP antagonists.

Sam Knight (§2.6.6). Only one region surrounding a target gene was positive for a large duplication: that of *BMPER*, in a single sample. I ran the structural variant caller Pindel on the corresponding BAM file. The coordinates of the purported duplication were chr7:33129823–33187428, as shown in figure 6.3a. Thirteen reads supported this call: ten with anchors upstream, and three with anchors downstream of the variant. This was an underwhelming amount of evidence, especially given the large distance of the variant from the gene of interest ($\sim$ 760kbp from *BMPER*).

Nonetheless, the potential duplication deserved follow-up. The individual in question (coded IT9) had additionally been typed on the OncoArray panel along with members of his family. To validate the call, I loaded these data into IGV and examined LogR ratio at the locus. Neither IT9 (figure 6.3b, top), nor his two sisters (figure 6.3b, middle and bottom) showed evidence of structural variation in the region.

6.2.2.1 ChIP-seq

Much of the current noncoding genome has been annotated thanks to the ENCODE consortium (The ENCODE Project Consortium et al., 2012). This expanding catalogue of chromatin states includes transcription factor binding sites in cells of multiple human tissues, including HCT-116 CRC cells. However, as the ENCODE consortium helped to show, transcription factor binding is responsive to cell type and stimulus (Neph et al.,

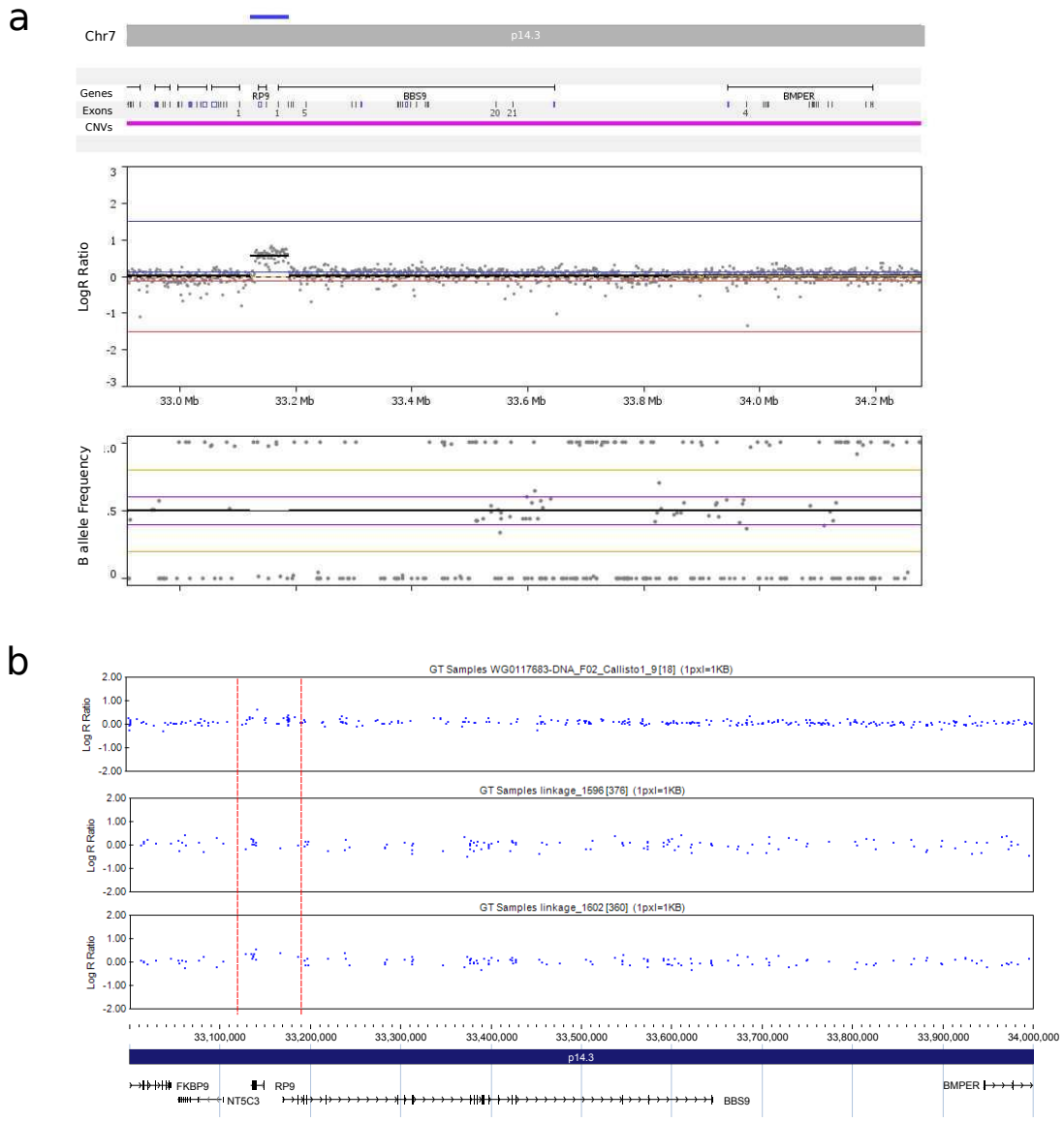


Figure 6.3: Putative duplication in the *BMPER* region. *a*, Nexus trace showing duplication (blue highlight) of a region upstream of *BMPER* in IT9; *b*, the same region (red lines) in OncoArray calls of IT9 (top) and two affected sisters (bottom two LogR traces) in IGV. No duplication is seen.

2012). No non-cancer colorectal cell ChIP-seq data is available through ENCODE. Nor are binding sites currently assessed following appropriate signalling activation.

I hypothesised that regions of colorectal BMP-responsive SMAD binding might be enriched for germline mutation within the Tomlinson set. I used the immortalised colonic epithelial cell line 1CT (§2.11.1), treating them with 100ng/ml recombinant human BMP2 or BMP4, or with BSA/HCl vector, for 16 hours. 1CT cells have an intact SMAD1/5/9 signalling pathway (figure 6.4a) that responds to this concentration of BMP2 and BMP4 (figure 6.4b–c). With the help of Chaido Stathopoulou, I provided Dr. James Platt with treated 1CT cells in triplicate for ChIP-seq. This pulled down the transcription factors at the promoters of known BMP-responsive genes such as *ID1* (figure 6.4d).

I used merged replicate T-PIC peaks calls to leniently define SMAD1/5 binding regions for each condition. Two findings were surprising. First, the ChIP-seq did not enrich for known SMAD motifs. Second, there was little overlap between genomic peak distribution between treatments (figure 6.5a). A peak was declared the same in two conditions if more than 50% of its span in either call was shared with the other, using BEDtools. The plurality of peak calls for each condition was private. Using the Standford University webtool GREAT, I saw that genes near BMP4-treated peaks alone were enriched for gene ontology terms that included negative nodal regulation (binomial $Q = 6.9 \times 10^{-5}$), and WNT receptor signalling involved in wound healing and epidermal spread (binomial $Q = 4.7 \times 10^{-5}$).

I binned rare single nucleotide variants (singletons, doubletons, and alleles with derived frequency of 0.5–1%, 1–5%, and over 5%) in the 275 Tomlinson genomes by presence within either control SMAD1/5 ChIP-seq peaks, or within BMP2- or BMP4-treated peaks that did not overlap with control calls. The BMP-activated bins were not enriched for rare variation (figure 6.5b, Kruskal-Wallis rank-sum $P = 0.99$; compare to figure 5.3d).

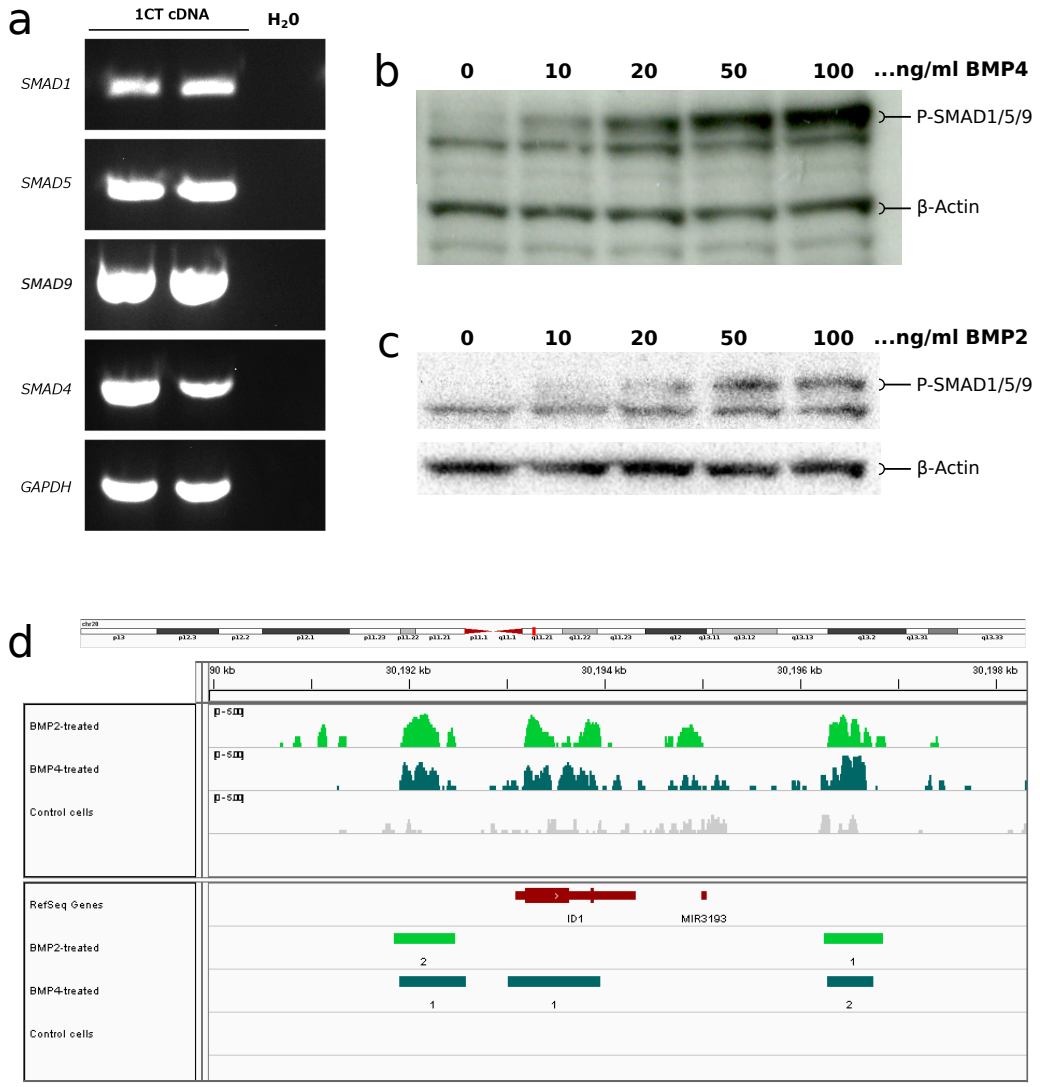


Figure 6.4: 1CT cells have an intact BMP signalling pathway for SMAD1/5 ChIP-seq. *a*, 1CT cells express BMP-specific SMADs by RT-PCR; *b* and *c*, 1CT cells phosphorylate SMAD1/5/9 in response to 16 hours' exposure to > 50ng/ml recombinant human BMP4 and BMP2, shown by representative Western blot; *d*, IGV-visualised ChIP-seq normalised reads (top) and T-PIC-called peaks (bottom) are present at the *ID1* promoter in BMP2- (light green) and BMP4-treated 1CT cells (dark green), but not in cells treated with vector (grey). Chromosome 20:30,190,000–30,198,000 is shown.

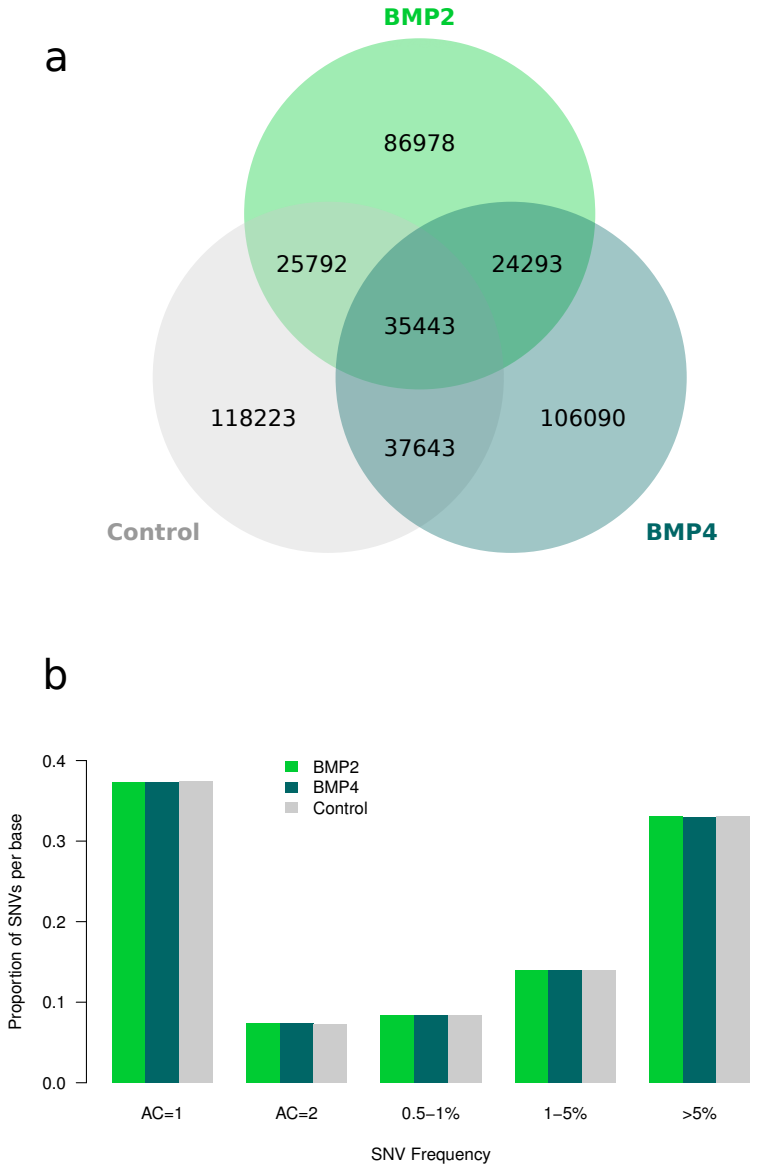


Figure 6.5: No enrichment of rare variants in SMAD1/5 binding sites. *a*, the number of called peaks in each treatment condition, with overlapping peaks assessed by BEDtools; *b*, single nucleotide variants per base of ChIP-seq peaks in BMP2-, BMP4-, and vector-treated 1CT cells, per variant frequency bin.

6.2.3 Patient subsets

I next searched for high-confidence loss of function variants present in three curated non-exclusive subgroups of the Tomlinson patient genomes. These were: those with CRC before the age of 45 ("early onset" group, $n = 63$); those with ten or more adenomas before age 40 ("multiple adenomas", $n = 18$); and those with ten or more serrated polyps before age 35 ("serrated polyposis", $n = 11$). This last definition deviates from the WHO criteria for SPS outlined in §6.1.2. As also argued by Gala et al. (2014), these criteria are unnecessarily restrictive for candidate study, requiring both more family information than is often available for recruited patients, and a more precise polyp categorisation than is provided by most pathology reports, sometimes because the reports predate modern serrated polyp subtypes. The age restriction I impose is a proxy for syndromic disease in the absence of widespread pedigree information.

I filtered the sample subgroup variants as follows. Variants were excluded if they lay within a superduplication alignment, REPMASK simple repeat sequence, area of Phastcons conservation score < 0.1, or LOFTEE "low confidence" annotation. Variants were required to be rarer than 0.1% in ESP6500 and present at allele count < 10 in the WGS500 control genomes. Finally, I removed any variant predicted to affect 98 loss of function-tolerant genes, as determined by MacArthur et al. (2012), and retained variants predicted by VEP to have a high probablility of functional consequence. The resulting variants, minus olfactory receptor gene variants, are displayed in table 6.4.

Neither the multiple adenoma nor the SPS subsets relinquished sensible pathogenic variants via this filtering strategy. *BAZ2A* is a gene involved in epigenetic silencing whose overexpression predicts prostate cancer recurrence (Gu et al., 2015); although not present in the WGS500 controls, its start-loss variant is seen 55 times in ExAC (a 0.11% frequency). The *IFT57* start-loss similarly is present 240 times in ExAC (0.21% frequency), including two homozygotes. IFT57, also known as HIPPI, is a pro-apoptosis caspase-8 interactor and transcription factor (Datta et al., 2011). These two genes are certainly involved in tumour-suppressive biology, but their loss of function variants are insufficiently rare in ExAC to be plausibly causal. The other SPS high-impact variant uncovered was in *KLHDC7B*, a poorly annotated gene which is not expressed in the

chr	Gene	Variant	Effect	Sample
1	*CHML*	rs763865507	TP332–333fs.	LP6005794-DNA_C03
3	*IFT57*	rs112224477	start lost	LP6005223-DNA_F01
4	*LNX1*	rs376599477	start lost	LP6005912-DNA_A03
6	*PGK2*	rs751994592	G347fs.	IT29
7	*MALSU1*	rs200950879	start lost	LP6005795-DNA_H11
7	*TRBV4-1*	rs56090737	Y54*	LP6005794-DNA_D09
				LP6005795-DNA_E10
				LP6005796-DNA_A03
12	*APOBEC1*	rs139646668	start lost	LP6005794-DNA_F08
17	*MFSD6L*	rs754195891	K427fs.	LP6005912-DNA_A02
17	*NAT9*	rs377529868	start lost	LP6005796-DNA_E02
20	*DEFB132*	rs188183503	start lost	LP6005796-DNA_H05
12	*BAZ2A*	rs201499595	start lost	IT25
3	*IFT57*	rs112224477	start lost	LP6005223-DNA_F01
22	*KLHDC7B*	rs746113253	R180fs.	LP6005223-DNA_D02

Table 6.4: High-likelihood functional variants within Tomlinson genome phenotypic subsets. Top, early onset samples; middle, multiple adenomas group; bottom, SPS samples. Note that the *IFT57* patient is both an early onset and an SPS patient. All variants are present heterozygously; fs., frameshift.

colorectum (Human Protein Atlas), and which has not previously been associated with disease. Although the Tomlinson frameshift mutation is only present twice in ExAC, its high-impact predicted function largely comes from the gene's single-exon structure, and it is not a good candidate for follow-up.

In the early-onset group, one "gene" was multiply deleteriously variant: the T cell receptor β variable 4-1 block (*TRBV4-1*) used for β chain creation by V(D)J recombination within the 7q34 T cell receptor cluster. It is difficult to imagine, however, how the heterozygous loss of one portion of the T cell receptor potential coding space might impact cancer risk. Moreover, the variant is present 80 times in ExAC (0.06% frequency), and once in the WGS500 controls.

Like the *TRBV4-1* stop-gain, two other candidate early-onset variants were present once in the 144 WGS500 controls: the *APOBEC1* start-loss (ExAC allele count = 187, 0.15%), and the *DEFB132* start-loss (ExAC allele count = 23, 0.02%). Not only are they reasonably frequent in ExAC, *DEFB132* is only expressed in the testes, as are *PGK2*, and *MFSD6L* (Human Protein Atlas). Another gene harbouring an only moderately rare variant, *NAT9*, is not expressed in the colon, and is poorly functionally annotated

(start-loss ExAC allele count = 35, 0.03%). The loss of function of all of these genes is unlikely to be pathogenic

Two of the three remaining early-onset loss of function variants are of mild interest. *MALSU1* is expressed in all tissues; it is required for mitochondrial ribosome assembly. Its start-loss is seen 36 times in ExAC (0.06% frequency). However, its impingement on cancer biology is difficult to discern. The *CHML* loss of function frameshift is rare (ExAC allele count=4, 0.003%), but again within an unpromising gene. The single-exon *CHML* is expressed in colon, where its product geranylgeranylates Rab proteins.

The *LNX1* start-loss is also rare. It is seen in ExAC 5 times, a frequency of 0.004%. It was not present in the WGS500 controls. *LNX1* encodes an E3 ubiquitin protein ligase which is expressed in the colon and which is downregulated in gliomas (Chen et al., 2005). Its heterozygous knockout might contribute to disease in `LP6005912-DNA_A03`, and merits further investigation. However, the main result from these patient subset searches remained that no full protein-coding gene suffered multiple losses of function within any group.

6.2.4 Linkage analysis

Some members of the Tomlinson genomes suffered familial CRC that appeared Mendelian, and were immediately related to individuals whose non-WGS genotypic information was available (§2.6.8). I looked more closely at two such pedigrees, coded OX2027 and OX70783.

6.2.4.1 OX2027

The index patient in this family is encoded IT9. This is the same proband whose *BMPER* duplication I ruled out in §6.2.2. IT9 has three sisters, all with a colorectal phenotype. IT9, who is male, presented at 50 with stage III rectal adenocarcinoma, which later metastasised to his liver. Two of his sisters were diagnosed with CRC, with the youngest subsequently dying of the disease, aged 44. The middle sister had two colorectal polyps removed at age 46 and another three removed at 52. The family's father was diagnosed with CRC at 64.

IT9 is married with three children. His daughter presented with colorectal polyps at 34, and his middle son had the same aged 25. IT9's youngest son has no reported colorectal pathology, although this is in the absence of formal investigation.

An abridged pedigree of OX2027 is shown in figure 6.6a. The inheritance pattern of CRC or polyposis strongly suggests a Mendelian, autosomal dominant predisposition trait. IT9's two deceased uncles both reportedly died of colorectal disease, but this has not been confirmed by medical records.

The individuals in pedigree OX2027 for whom I had genetic information were IT9 and his three sisters (see §2.6.8). Nonparametric linkage analysis performed on these four individuals revealed an area of chromosome 12q that cosegregated with disease, with the minimum possible exponential K&C P value for this dataset, $P = 0.002$. Nonetheless, $LOD < 3.0$. See figure 6.6b.

There was no detectable large copy number change present in the IT9 genome under the LOD peak by B-allele frequency (§2.6.6). Nor were there smaller variants that were both rare and obviously functional. I removed alleles under the peak (approximately chr12:106,000,000–110,000,000) which were present more than twice in WGS500 controls, leaving 17 variants. None were coding, and only one passed the "low" VEP threshold for effect in any transcript. This was an upstream variant of *DAO* (D-Amino Acid Oxidase), rs185813301, which alters a splice site in some transcripts. *DAO* is not expressed in the human colon (Human Protein Atlas), and rs185813301 was unconvincing as a pathogenic mutation, despite its rarity (its minor allele is detected 4 times in ExAC, a frequency of 0.008%).

I therefore used the Broad Institute tool Haploreg v4.1 to assess function of the 17 rare variants. Two SNVs in particular had reasonable regulatory potential. These were rs61941589 (chr12:109,914,856:T>C) and rs28384416 (chr12:109,915,183:C>G) within *KCTD10* (potassium channel tetramerisation domain containing 10). Both were in ChromHMM transcription start site states (state 1 in the core 15-state model, state 3 in the 25-state model) in colon and rectal samples, and were in gastrointestinally open chromatin in DNase assays (Ward and Kellis, 2012).

rs61941589 is within the first intron of *KCTD10*. It alters both SMAD4 and SMAD1

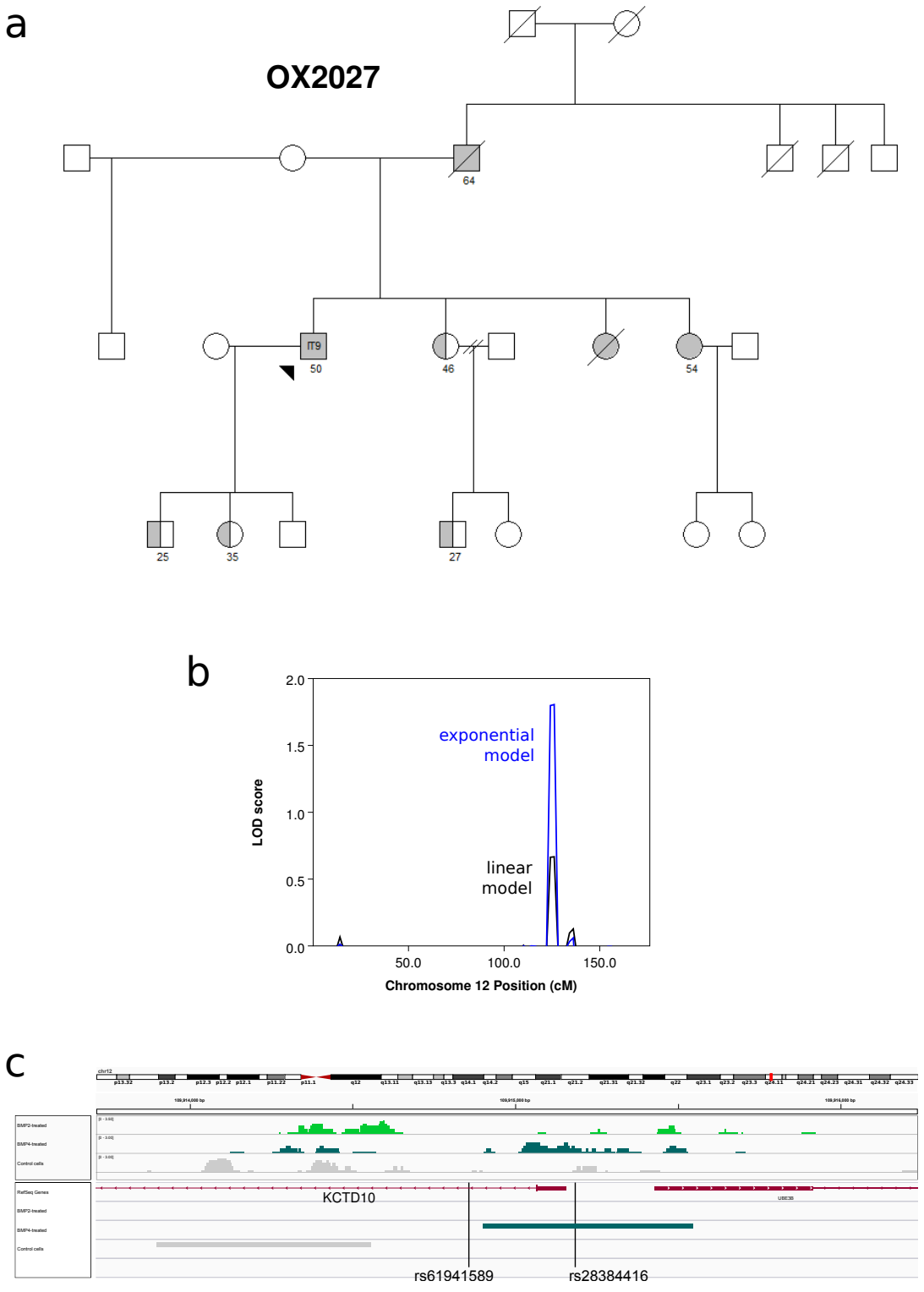


Figure 6.6: *a*, simplified pedigree of family OX2027. Half-shading indicates polyposis; full shading indicates CRC. The number under shaded boxes corresponds to age at presentation of the relevant phenotype, if known. IT9 is indicated by label and arrow. *b*, the *LOD* peak at chromosome 12 in family OX2027; *c*, SMAD1/5 ChIP-seq peaks at the *KCTD10* promoter region.

motifs, as catalogued by ENCODE (Kheradpour and Kellis, 2014), although it is not under a ChIP-seq peak in BMP-treated 1CT cells (figure 6.6c). Moreover, the SNV is an eQTL for *KCTD10* expression in the GTEx database (The GTEx Consortium et al., 2015): in transformed fibroblasts ($P = 1.17 \times 10^{-7}$) and in subcutaneous adipose tissue ($P = 8.28 \times 10^{-6}$). The minor allele is present once in WGS500 controls, but is seen 568 times in 1000G phase 1 (11%, although this frequency falls to 1% in Europeans).

rs28384416 is a 5′-UTR variant that lies directly within the *KCTD10* CpG island (chr12:109,915,132–109,915,638; Observed CpG/Expected CpG = 0.81). I saw no variant allele in WGS500 controls; the minor allele is present in 1000G phase 1 at a 3.4% frequency, which is lower in Europeans. The base is conserved by SiPhy-omega (Garber et al., 2009), and its alteration disrupts six TRANSFAC motifs (MAF, REST, NKX2-9, SPI1, RAD21, and SREBP-1). It lies within a T-PIC-called SMAD1/5 peak in BMP4-treated 1CT cells (figure 6.6c). Moreover, the SNP is bound by 14 transcription factors in ENCODE, including RNA polymerase II in colon cancer HCT-116 cells. It modifies the promoter of an active colorectal gene.

6.2.4.2 OX70783

The OX70783 proband, a 74 year-old male coded `LP6005223-DNA_D02` and referred to henceforth as D02, has had at least seventeen hyperplastic polyps removed on colonoscopy, and has suffered other cancer types (one pituitary adenoma and one basal cell carcinoma). The brother of D02 developed CRC aged 39. A simplified pedigree of OX70783 is shown in figure 6.7a. The individuals in this pedigree for whom I had genetic information were D02 and the four oldest children of his brother (figure 6.7a, purple outline; see §2.6.8).

The extended pedigree around D02 displays a degree of consanguineity: D02's parents are second cousins. Some individuals are of a small Canadian ethnic minority called White Métis. Both inbreeding and small recent population size are risk factors for recessive genetic disease, but the inheritance pattern of the OX70783 familial serrated polyposis is dominant.

The SPS affection status of OX70783 members was often ambiguous due to incom-

plete pathological information. I did not have size or location information for most excised polyps and could not therefore conform to WHO specifications as laid out in §6.1.2. Especially, to diagnose based on a first-degree relative with SPS (criterion *ii.*) would have been unhelpful. I therefore used a lenient version of criterion *i.*, which is to say ≥ 5 serrated polyps.

This threshold named D02 as affected, along with his eldest niece and nephew. The eldest niece had been whole-genome sequenced: `LP6005223-DNA_H03` (table 5.2; referred to henceforth as H03). H03 is 48 and has had 11 sessile serrated and 3 pedunculated polyps removed on colonscopy. The nephew is 41 and has had 14 serrated polyps. The two other nieces whose DNA was available, N1 (aged 39) and N2 (aged 37), have had under 4 and 2 polyps, respectively, with no histopathology available. OX70783 members undergo biennial colonoscopic screening. The two polyps of N2 were excised at age 26; she has had no further colorectal pathology since.

Because I had access to two whole-genome sequences, and because minimum exponential K&C P value was larger than for OX2027 (maximum nonparametric $LOD = 0.6$), I took two parallel approaches to searching for pathogenic variation in OX70783. One, I took the intersection of the two OX70783 genomes, regardless of linkage, to retain heterozygous shared variants predicted to have "high" functional effect likelihood by VEP that were not present even once in WGS500 controls. Two, I performed linkage analysis on the family and looked at the intersection of the two whole genomes at regions of nominal exponential K&C significance ($P \leq 0.05$), filtering on WGS500 control allele count, which I required to be below 2, and ExAC allele count, which I required to be below 100. Both methods converged on one candidate.

The first approach left the seven variants in table 6.5. Some of these variants are not rare in other datasets: the *NBPF1* indels are present at 5025 times in ExAC (rs373748839) and 1912 times in ESP6500 (rs368729374), and the *QPCT* stop-gain is present 64 times in ExAC.

The *KLHDC7B* indel was reported earlier in this chapter, in table 6.4, and is only seen twice in ExAC. As I previously stated, the gene is poorly annotated, is not expressed in colon, and has not been associated with disease. The *MAN2B2* stop-gain is seen only

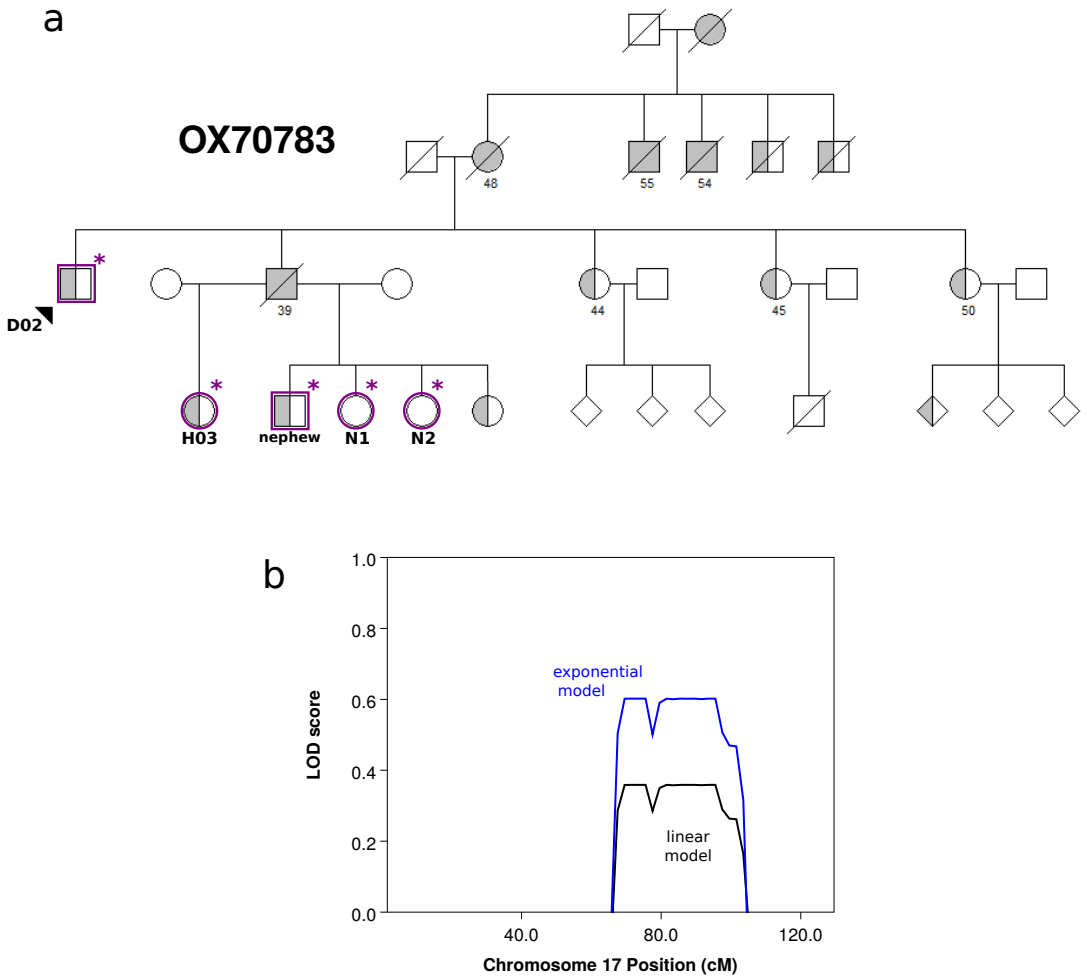


Figure 6.7: *a*, an abridged pedigree of family OX70783. Half-shading indicates serrated polyposis; full shading indicates CRC. The number under shaded boxes corresponds to age at presentation of the relevant phenotype, if known. The proband, D02, is indicated by the black arrow. Purple outlines and asterisks mark genotyped individuals. *b*, the *LOD* peak on chromosome 17 overlying *RNF43*.

chr	Gene	Variant	Effect
1	*NBPF1*	rs373748839	frameshift
1	*NBPF1*	rs368729374	frameshift
2	*QPCT*	rs142847125	R133*
4	*MAN2B2*	rs759467902	Q354*
15	*AP3B2*	83332566:C>T	W789*
17	*RNF43*	56440746:TC>T	frameshift
22	*KLHDC7B*	rs746113253	frameshift

Table 6.5: Rare, high-effect variants shared by the two OX70783 genomes. All variants are heterozygous in both uncle and niece, and not present in 144 WGS500 controls.

once in ExAC, but again is a gene of uncertain significance. The W789* stop-gain in *AP3B2* has not elsewhere been reported. However, the gene's expression is confined to the central nervous and endocrine systems.

The second method revealed multiple regions across the genome with exponential K&C $P \leq 0.05$, with no single peak as in OX2027. See, for example, the chromosome 17 linkage landscape in figure 6.7b. Eleven shared variants under *LOD* peaks were predicted to have at least moderate impact by VEP, including variants in *GANC*, *TBC1D2B*, *SOS2*, *PKD1*, *CD40*, *GARS*, *NCOA3*, and *SFTPA2*. Two moderate-effect missense variants under *LOD* peaks were in striking genes. Neither was present in ExAC or in WGS500 controls. The first was a K115E change in exon 4 of *BRCA2* (rs56242644). Mutation of *BRCA2* is associated with increased risk of breast and ovarian cancer (Antoniou et al., 2006; King et al., 2003). The second was P333A in *KREMEN2*, a DKK1 receptor which blocks WNT signal by causing LRP6 endocytosis (Mao et al., 2002). Three *LRP6* mutations were implicated in CRC pathogenesis by de Voer et al. (2016). However, neither missense was as likely an explanation of disease as the only *LOD*-overlaid variant given "high" impact prediction by VEP: the undiscussed last variant in table 6.5, chr17:56440746:TC>T.

This *RNF43* frameshift replaces leucine 157 with the amino acid string LPGQWC*, a premature STOP in exon 5 of 10. The variant is not present in ExAC. *RNF43* has been associated with SPS twice before, although disease cosegregation was not proved (Gala et al., 2014; Taupin et al., 2015). I confirmed the presence of the mutation in D02 and H03 by Sanger sequencing, and showed that it cosegregated with my definition of SPS in OX70783 (figure 6.8a). This is consistent with its presence in a block of maximal K&C exponential *LOD* ($P = 0.05$) as shown in figure 6.7b. The OX70783 *RNF43* frameshift occurs at a conserved region of the gene (figure 6.8b) within the same extracellular domain harbouring both previously reported pathogenic changes (figure 6.8c).

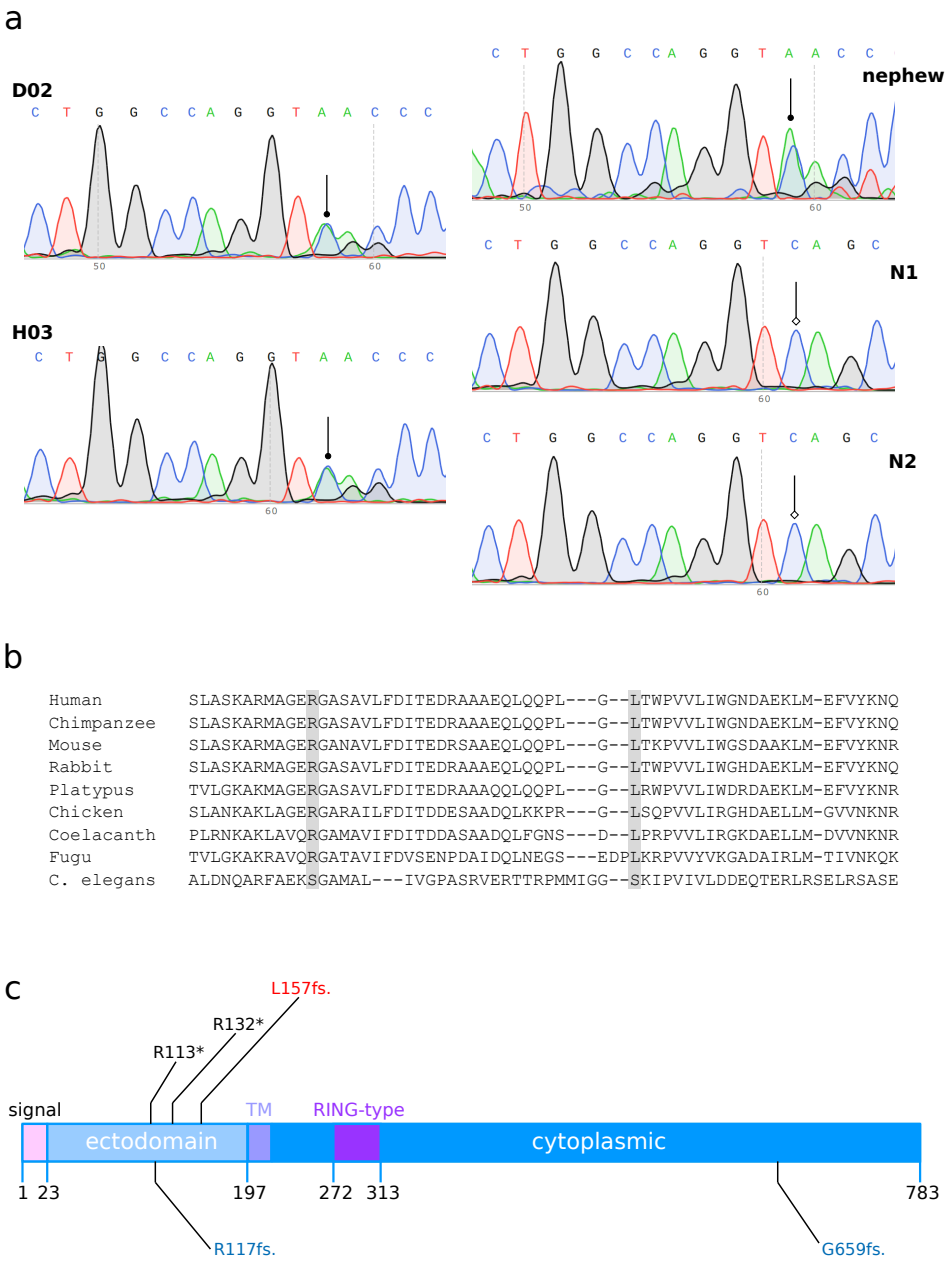


Figure 6.8: Cosegregation of mutant *RNF43*. *a*, Sanger sequence traces of OX70783 family members, with frameshift location indicated; *b*, conservation of amino acid sequence in positions 121–180 from human to worm BLAST alignments, highlighting the Taupin et al. (2015) arginine 132 (first grey region) and the position of the OX70783 leucine 157 frameshift (second grey region); *c*, simplified structure of RNF43, showing signal peptide, PA-associated ectodomain, helical transmembrane section (TM), and the RING-type finger motif within the cytoplasmic portion of the protein. Below in blue are the two most common somatic alterations in colorectal and endometrial cancers (Giannakis et al., 2014); above in black are the positions of the previously reported germline changes at arginines 113 (Gala et al., 2014) and 132 (Taupin et al., 2015) claimed to cause SPS; in red is the position of the OX70783 mutation.

6.3 Discussion

I have reported a number of germline genetic changes that may contribute to CRC pathogenesis. These include two novel *MLH1* variants, a *POLD3* frameshift, *EXO1* and *RPA4* stop-gains, and an *LNX1* start loss.

Three aspects of this chapter bear more thorough discussion. First, I found a new instance of *MSH3* homozygous CRC pathogenicity. Second, I reported two areas of significant linkage with familial CRC. In one I could not precisely name the underlying causative variant; in the other I found a new instance of *RNF43* SPS association, for the first time providing segregation analysis of the mutation. Third, I made tentative steps into the noncoding genome. I will elaborate on each below.

The *MSH3* proband is particularly interesting in the context of four other conditions. The first is naturally the previous *MSH3* mutation. The individual in this thesis has a considerably milder phenotype than that reported by Adam et al. (2016), whose two patients displayed thyroid, breast, uterine, and brain tumours, as well as polyposis and CRC. By contrast, my proband had normal age of onset for CRC (62), followed by a much milder polyposis. Although her adenocarcinoma was EMAST-positive, it was stable at mononucleotides. This is not so much of a contradiction as it first appears: MSH3 does not process mononucleotide lesions, and Adam et al. (2016) also saw mononucleotide stability in their *MSH3* mutant adenomas. Dinucleotide and tetranucleotide stability is more important in this context. Haugen et al. (2008) reported a series of 117 sporadic CRC tumours where 35 of 82 (42.6%) MSS cases exhibited EMAST. They suggested that, rather than being truly stable at all microsatellites, MSS/EMAST tumours are low-level MSI at non-tetranucleotide sequences; this unearths an older debate about the accuracy of a strict MSS/MSI binary (Halford et al., 2002). It also highlights an imperfect practice of hospital genetics laboratories: mononucleotides alone are used to assess MSI status, because paired tumour and normal samples are not required. This may be inappropriate.

The next two conditions to consider are Lynch syndrome, which is by definition a dominant disease, and constitutional mismatch repair deficiency syndrome (CMMRD),

which is recessive. Turcot et al. (1959) described a condition of primary brain tumour combined with colorectal adenomas; "Turcot's syndrome", considered too restrictive a diagnosis (Østergaard et al., 2005), has subsequently been folded into CMMRD, where biallelic germline mutation of the Lynch genes, *MSH2*, *MLH1*, *MSH6*, and *PMS2*, causes a mixture of haematological cancers (principally non-Hodgkin T lymphomas), brain tumours (mostly high-grade gliomas), colorectal polyps and tumours, and other malignancies.

CMMRD is phenotypically different to Lynch syndrome. Blood and brain cancers are more often the first to present (Lavoine et al., 2015; Wimmer and Kratz, 2010). When it does occur, CRC in CMMRD is more often left-sided (Durno et al., 2015; Lavoine et al., 2015), and is not necessarily MSI (Will et al., 2007). CMMRD caused by *MSH2* and *MLH1* homozygous mutation in particular rarely manifests Lynch-like cancers (Lavoine et al., 2015). However, in some cases the overlap is clear. Gallinger et al. (2004) described two children with homozygous *MLH1* mutations who developed duodenal and colorectal malignancies as early as 9 and 11. Will et al. (2007) reported a homozygous 400kb deletion of *PMS2* which also caused a strong gastrointestinal phenotype: 10 separate carcinomas were removed as part of the proband's panproctocolectomy at age 23. These are striking examples of CMMRD presenting with Lynch-like pathology. Additionally, although often secondary to other cancers, many CMMRD cases do develop Lynch-like tumours: in 92 CMMRD patients studied by Wimmer and Kratz (2010), 51 Lynch-associated tumours arose. Most exciting is the *PMS2* founder mutation found in the Quebecois Inuit of Nunavik (Li et al., 2015). This is an incompletely molecularly penetrant *de novo PMS2* splice site. Homozygotes develop Lynch-type early-onset cancers (astrocytoma, duodenal, stomach, and colorectal cancer), but retain residual *PMS2* expression. In other words, a molecular phenotype between that of Lynch syndrome and CMMRD produces a gross phenotype that also lies between the two.

If it is sensible to think of Lynch-like syndrome (Rodríguez-Soler et al., 2013), Lynch syndrome, and CMMRD as lying on a spectrum of phenotypic severity — which is contentious — then it's tempting to speculate that *MSH3*-related AR polyposis might lie on the same spectrum. This would fit with previous reports of *MSH3* mutation in

endometrial cancer (Risinger et al., 1996). Possibly because of the cachet of a "new" syndrome, Adam et al. (2016) were coy about this possibility, admitting only "some degree of overlap with CMMRD". The *MSH3* proband in this thesis suggests that this overlap would be at the milder end, milder still than Lynch syndrome. But this would not be the case if the hypothesis of Adam et al. (2016) bore fruit: that the EMAST instability seen in other tumour types should also be ascribed to MSH3 dysfunction, and possibly to germline *MSH3* mutation.

The fourth pertinent condition stretches this speculation further. MAP was discovered by Al-Tassan et al. (2002), and is a recessive deficit in the DNA base excision repair pathway. As well as polyps and CRC, MAP has been reported to produce the mild extracolonic phenotype of pilomatricomas, recalling CMMRD skin macules (Baglioni et al., 2005). More importantly, *MUTYH* is within the MMR KEGG ontology set. MUTYH and MSH6 physically and functionally interact (Gu et al., 2002), so *MUTYH* influences both break excision and mismatch repair. Indeed, Lynch-like syndrome is enriched for undiagnosed MAP (Castillejo et al., 2014), and there is some evidence that *MUTYH* and *MSH6* alleles can interact to affect CRC risk (Niessen et al., 2006), although more recent reports dispute this (Win et al., 2015). My suggestion is that the different degree of impingement on MMR of these different conditions might inform their severity and spectrum of phenotypes.

This may be an unhelpful degree of taxonomic fastidiousness. Indeed, the categorical confusions encircling Lynch syndrome are reminiscent of the blurred boundaries between FAP, Gardner's syndrome, and other polyposes as they were first being described, or of the disagreements about what is and isn't HMPS (see §1.2.4). The conclusion I wish to draw is not a worked-out interrelationship of MMR pathologies, but rather a petition for bottom-up disease classification. Instead of a focus on phenotype, or even on mutant gene, it seems sensible to move towards an allele-based description of this whole family of CRC syndromes. Within-gene variant alleles themselves should become pathognomonic, which ironically might make it easier to spot patterns across gene set variants. One immediate corollary is that the InSiGHT database of Lynch-causing alleles should expand to include reports of CMMRD and *MSH3* variants.

The particular *MSH3* stop-gain I report requires further research. It is important to see the effect of the mutation on RNA and protein expression in transfected cells; I hypothesise that the 5 exons preceding the premature STOP are not translated, leading to no tagged expression on Western blot. Moreover, the introduction of the mutation by CRISPR-Cas9 into MSS CRC cell lines should induce EMAST. Relatedly, I would also like to assess dinucleotide microsatellite stability in the *MSH3*-mutant tumour sample, which I would predict to be unstable.

I will now briefly consider the results of my linkage analyses. Although among the largest in the Tomlinson genomes, neither pedigree studied was sufficiently genotyped to power linkage discovery at traditional levels of significance ($LOD > 3$). Instead, in the manner of Palles et al. (2013), I used shared or segregating regions to narrow down whole-genome data.

This approach yielded a cosegregating frameshift mutation in *RNF43* in the SPS family, OX70783. Like LNX1, RNF43 is an E3 ubiquitin-protein ligase. Knockout of *Rnf43* and its sister gene *Znrf3* together caused rapid adenoma formation in mice, accompanied by β-catenin and Wnt target gene upregulation (Koo et al., 2012). HCT-116 CRC cells are *RNF43*-mutant and normally WNT3A-responsive; the reintroduction of transfected *RNF43* abolished this responsiveness, and downregulated frizzled receptors FZD1–3. RNF43 ubiquitinates frizzled WNT receptors, targeting them for degradation (Koo et al., 2012). It competes with R-spondins to set the level of WNT responsiveness in intestinal stem cells (Yan et al., 2017).

Gala et al. (2014) first connected *RNF43* to SPS, finding two arginine 113 stop-gains within 20 exome sequences by searching a set of 555 GWAS-related genes for strong loss of function mutations. They were lucky to pick up the mutation by this method, because their "GWAS" gene set was derived from the NHGRI-EBI GWAS catalogue. *RNF43*'s presence in the catalogue is unearned: it was leniently included based on a table of "suggestive" associations within a small Japanese pancreatic cancer GWAS, in which one missense mutation in *RNF43* (rs2257205) was associated with disease at $P = 1.6 \times 10^{-5}$ (Low et al., 2010). The *RNF43* locus has in fact not been associated at more reasonable P thresholds with any cancer by GWAS.

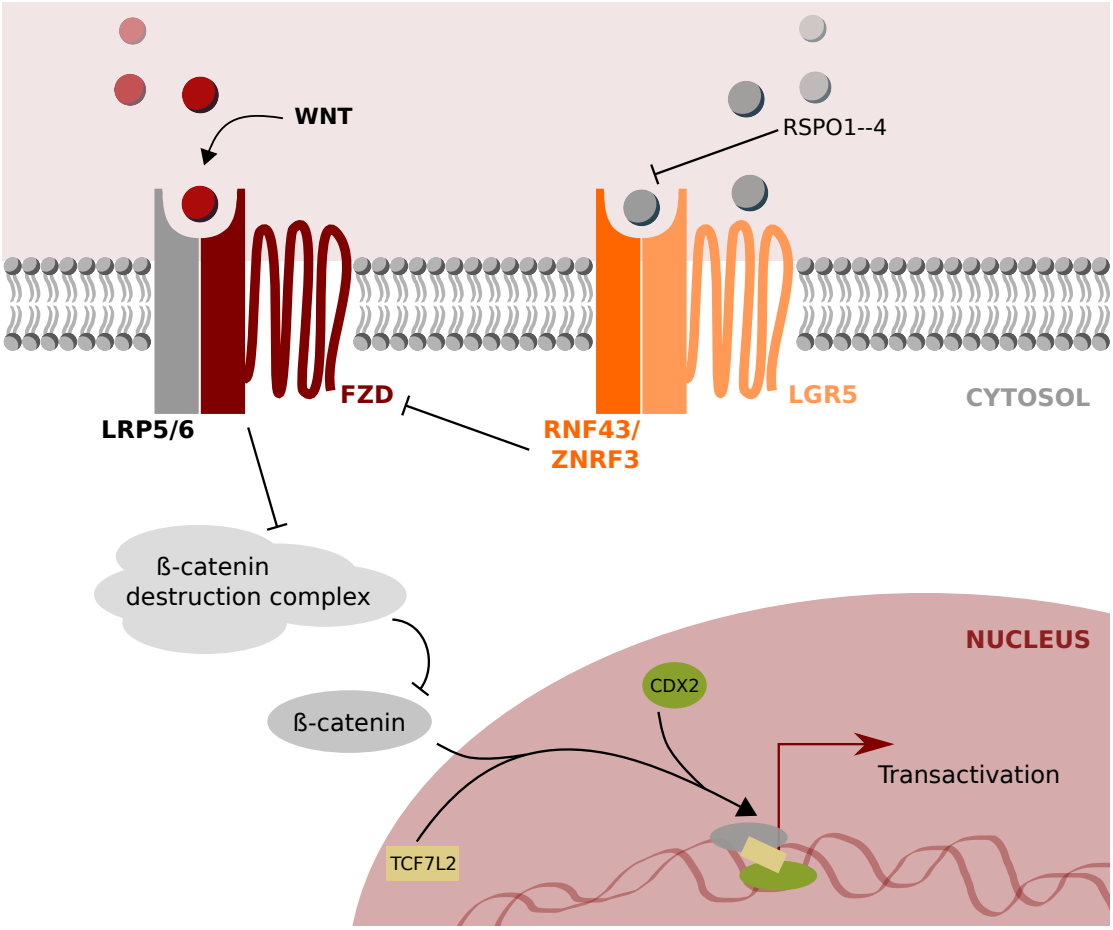


Figure 6.9: Simplified diagram of the WNT pathway. RNF43 and R-spondins compete to set the level of WNT responsiveness in intestinal LGR5$^+$ stem cells (Yan et al., 2017). The β-catenin destruction complex contains axin, APC, GSK3β. Example β-catenin cofactors are shown; many others are omitted.

The gene is nonetheless important in CRC. It was recurrently mutant in nearly a fifth of both TCGA and private CRC tumours sequenced by Giannakis et al. (2014). This was also the case in endometrial cancer. Most mutation of *RNF43* in both tumour types is at glycine 659 and arginine 117; see figure 6.8c. Giannakis et al. (2014) also determined that *RNF43* truncation conferred a fitness advantage to tumour cells, was associated with MSI, and was mutually exclusive with *APC* mutation. The frequent, functional, somatic mutation of *RNF43* is a better testament to the significance of inherited changes in the gene than its spurious GWAS evidence above. It's not clear how many Mendelian candidate genes remain to be found, but somatic driver mutations in those same genes in sporadic CRC would substantiate any future discoveries. Such is the long reach of the Knudson hypothesis, as discussed in §1.2.4.

The leucine 157 frameshift in OX70783 needs considerable molecular corroboration, especially because of the difficulty of accurately diagnosing SPS based on incomplete pathology reports with unsatisfying WHO criteria. I am both seeking an RNA sample from D02 and formalin-fixed paraffin-embedded tissue from the father of H03 to assess *RNF43* expression. I presume that the effect of the mutation is that the premature stop codon produces either a useless mRNA or residual protein, which are subsequently degraded. Suggestions of this hypothesis include a reduction in overall RNF43 expression on Western blot, or an allele-specific reduction in mRNA expression. This could be tested in D02 tissue, and by transfecting cell lines with mutant and wild-type alleles. Alternatively, the mutant protein product might persist, but function improperly. The behaviour of transfected cell lines compared to wild-type *RNF43* overexpressors would be informative in this context: for example, it would be interesting to see whether the OX70783 frameshift is incapable of abolishing WNT3A response in the *RNF43*-null HCT116. Ideally, I would also source tumour or serrated polyp material from the family to search for loss of heterozygosity at the mutation. These many future experiments notwithstanding, the OX70783 frameshift appears to provide the first multi-generational genetic account of SPS.

The result of linkage study in pedigree OX2027 was less clear. I have suggested that changes to the *KCTD10* promoter might underlie disease. This gene is expressed throughout the body, including in the large bowel. KCTD10 interacts with basal body proteins at the cilium transition zone (Chih et al., 2011). More relevantly, Zhou et al. (2005) discovered its rat orthologue by sequence similarity to two Polδ-interacting proteins. In 3T3 cells, overexpressed HA-tagged KCTD10 pulled down myc-PCNA if its C-terminal PCNA binding motif was intact (Zhou et al., 2005). The protein also precipitated myc-POLD2, the small subunit of Polδ. Both Polδ (Longley et al., 1997) and PCNA (Goellner et al., 2014) are involved in MMR.

However, the *KCTD10* promoter changes in IT9 are only a suggestion; the surer finding is the *LOD* peak at the region. *LOD* without certain cause is not an unusual result. In 103 CRC pedigrees, Papaemmanuil et al. (2008) found a maximal nonparametric *LOD* of 3.01 at 3q22, with the best evidence for linkage from families segregating

rectal CRC. After manual Sanger sequencing of 30 genes within the *LOD* peak, Papaemmanuil et al. did not identify a causal variant. They hypothesised, much as I am forced to in OX2027, that this variant or variants lay in the unannotated and therefore more intractable noncoding genome.

ENCODE made notoriously over-grand claims about the improvement in noncoding DNA annotation its results provided (Eddy, 2012). But the efforts of the consortium are nevertheless necessary for understanding intergenic variants, as are experiments like them such as my SMAD1/5 ChIP-seq. My brief account in this chapter of the preliminary results from this pulldown suggests two things. First, in the great variability of stimulus-dependent captured binding site, I reinforce the point I made in §4.3: that the delicacy of transcription factor binding architecture makes strong negative conclusions difficult. It's possible that CRC-predisposing germline variation is not enriched at SMAD1/5 binding sites, as my data suggest; it's also possible that I have looked in the wrong cell type, with the wrong dose of BMP, at the wrong time. The only way to avoid this sort of special pleading is the furtherment of ENCODE and ENCODE-like projects. Second, this furtherment is an essential requirement for pathogenic rare variant searches beyond the exome, which will need to group variants by meaningful functional categories. I return to this idea in chapter 7.

Chapter 7

Discussion

"From the nature of the case the theory is difficult if not impossible to analyze by direct experiment, and for this reason, as well as for its impracticability, it is probable that the hypothesis will not be favorably received by the medical profession."

Calkins (1914), reviewing the chromosome theory of Boveri (1914).

This thesis explores the inherited risk of colorectal cancer. Its contents reflect its timing: the work I have reported was performed after the GWAS boom of the mid 2000's, and at the start of the large cancer sequencing projects now being undertaken. As such, my work begins with post-GWAS fine-mapping, and ends with exploratory and candidate-based germline whole genome sequencing. This discussion is an opportunity to explore this transition, and the relationship between older and newer genetics modalities. In particular, whole genome sequencing is often painted as a fix to the disappointments and flaws of GWAS, its antecedent technology. I will briefly recap each chapter's work, and then will comment on this narrative. I will argue three things. First, that sequencing is not an upgrade of GWAS: they are complementary approaches that in some instances can resemble each other, but in other conformations can imply quite different models of disease. Second, many of the perceived failures of GWAS are nothing of the sort. This requires an exploration of the untapped missing heritability that is GWAS's greatest alleged crime. And third, the road from genetics to biology to bedside is long; some misconceptions about medical genetics might underlie the enthusiasm for sequencing; and attention should turn to more precise cancer GWAS phenotyping and high-throughput predisposition variant functional explanation.

In chapter 3, I reported a novel CRC risk SNP at the *GREM1* locus. I discovered this signal, at rs17816465, by stepwise regression, and then validated it to genome-wide significance in a collection of replication cohorts. rs17816465 lies within a *FMN1* intron, downstream of *GREM1*; I could provide no sure links between the SNP and the gene, either by conformation capture or eQTL. Nor was the SNP furnished with active chromatin marks in *GREM1*-expressing cell lines. However, rs17816465 displayed allele-specific luciferase activity, and elements within *FMN1* introns have been shown to regulate *GREM1* (Zúñiga et al., 2012). I also further explored the original tag's signal, creating a local four-SNP model of CRC risk which probably incorporates a haplotypic signal at the *GREM1* promoter. Every component SNP in this model has independently reached genome-wide significance. As I mentioned in chapter 3's discussion, my improved account of CRC risk at the *GREM1* locus is the second refinement of the 15q13 GWAS signal. The first, by Tomlinson 2011, has been consistently ignored by the subsequent literature, reflecting an irritating elision in fine-mapping papers: an ambiguity about whether a better tag or a better causal candidate is sought. Haplotypic or multi-SNP models are often considered only superficially.

In chapter 4, I complemented my fine-mapping at *GREM1* with additional examination of the signal at *POLD3*. I discovered overwhelming LD, which belies previous fine-mapping refinements. The only sensible account of my data was a large risk haplotype at *POLD3*, making the region fairly intractable to further genetic analysis. I asked whether a tag of this haplotype, or *GREM1* risk SNPs, might affect outcome after tumourigenesis. No CRC risk SNP alone was associated with patient disease-free survival time, though *GREM1* expression was an unfavourable prognosticator. Indeed, a combined score of all independent CRC GWAS risk SNPs also failed to predict survival in VICTOR and QUASAR2. However, this score did show a significant correlation with the degree of tumour driver gene mutation. The possibility is that GWAS SNPs bring a precursor closer to some neoplastic threshold, reducing the precursor's need for driver mutation.

In chapter 5, I remapped and recalled 299 whole genomes, which I combined with 196 older samples in order to explore allele frequency distributions and potentially in-

Quality	*GREM1*	*POLD3*
Link to CRC syndrome	HMPS	PPAP
Somatically mutated	no	yes
GWAS signal	yes	yes
Fine-mapping possible	yes	no
Open chromatin	partially	-
"Functional" DNA	at rs16969681	-
Fine-mapping to colon eQTL	at promoter	possible
Variant loops to promoter	no	-
Luciferase assay positive	yes	-

Table 7.1: What makes a good sporadic CRC effector candidate? A summary of evidence provided by this thesis at two loci in favour of their influence on colorectal tumourigenesis. "Functional" DNA means transcription factor motif or binding seen by ChIP.

formative population structure. I found that GATK was a superior multi-sample caller compared to Platypus, and that singleton variants were more numerous in foreign samples, especially in those with African heritage. I picked up 24 individuals with undetected or noncanonical mutations in known polyposis syndrome genes, and described the incidence of non-CRC Mendelian alleles, both dominant and recessive, in the genomes. Principal component analysis revealed little British population structure, once foreign individuals had been removed. However, rare variant sharing showed a geographic cline.

In chapter 6, I used a number of filtering and candidate-based methods to search for novel germline colorectal cancer predisposition genes. The two most interesting findings here were an individual homozygous for MSH3 loss of function and a new *RNF43*-mutant SPS pedigree. *MSH3* is a mismatch repair gene whose definitive involvement in Lynch-like CRC causation has been reported just once before (Adam et al., 2016). *RNF43* has twice been attached to SPS, but never with convincing cosegregation. With some categorical uncertainty, this thesis is the first report of such cosegregation.

All my reported work shares an interest in germline influences on CRC tumourigenesis. The evidence it has provided in support of various candidate genes is summarised in tables 7.1 and 7.2. I will use this discussion as an opportunity to resist an easy account of this thesis. My work was emphatically not a movement away from a frustrated model of germline cancer genetics towards modern sequencing efforts. Instead, I want to provide a brief defence of the GWAS model of complex disease, to explore the potential of whole genome sequencing, and to comment on the clinical utility of germline research.

Quality	*MSH3*	*RNF43*
Rare variant	yes	yes
Sanger validated	yes	yes
Strong functional impact predicted	yes	yes
Strong functional impact shown	EMAST	no
CRC pathway affected	MMR	WNT
In linkage peak	no	yes
Cosegregates	no	yes

Table 7.2: What makes a good familial mutation candidate? A summary of evidence provided by this thesis at two loci in favour of their influence on Mendelian disease.

This medical aim is often assumed to be the erection of a complete and predictive model of complex disease risk. As I found in chapter 4, this sort of predictive power is beyond current GWAS results, not simply due to unknown environmental factors. In short, there is missing heritability.

Missing heritability can be all things to all researchers (see in particular the various positions of Eichler et al., 2010). For epigeneticists, it is a sign that a significant contribution to trait biology comes from the epigenome; for exome sequencers, it implies that rare variants of large effect remain undiscovered; for whole genome sequencers, it speaks to the untapped dark matter of the noncoding genome, and rare variants therein; and for holdout GWAS proponents such as this author, missing heritability most strongly supports an infinitesimal, long tail of small effect size, common variants, which are below current GWAS significance thresholds for want of detection power.

An epigenetic plug for the heritability gap is a particularly strange suggestion, given the explicit absence of epigenetics from estimates of heritability. Narrow-sense heritability (h_2) calculations disregard the environmental input of epigenetics, so it's hard to see how epigenetics can account for unmeasured narrow-sense heritability. Doubtless DNA methylation and other mechanisms have some role to play in explaining phenotypic variance, but they cannot fill holes in models that never expected them. It's as irrelevant to the issue at hand as pointing out that a good deal of phenotype will be determined by an invariant genetic substrate that will be invisible to GWAS. Both points are true — trivially true, even — but appear not to appreciate the genesis of a narrow-sense heritability gap, nor to engage sensibly with the problem on its own terms.

Superficially an additional suggestion makes almost the same mistake, which is to incorporate interactive terms into heritability estimates. At first glance this solution is as irrelevant as the last, since interactive terms are not part of standard narrow-sense estimates (see §1.3, and in particular equation 1.2). However, environmental components are explicitly accounted for in the ACE estimates of h_2, whereas the additivity of genetic input is *implicitly* assumed. Zuk et al. (2012) argue convincingly that this assumption in fact inflates estimates of h_2, creating a "phantom heritability" that will perpetually frustrate researchers' efforts. However, this position is mildly contentious. Not only is the degree of trait epistasis difficult to measure (as admitted by Zuk et al., 2012), but monozygous correlation coefficients are almost always double those of dizygotes, as overwhelming additivity would predict (Polderman et al., 2015). In any case, the slight inflation of the total h_2 estimate, the target to be reached, is unlikely to account entirely for missing heritability.

This leaves us with two non-exclusive possibilities. Either missing variation is common, or it is rare.

If it is common, why hasn't GWAS seen it? As I illustrated in figure 3.3, GWAS loses power to detect risk association as effect sizes and allele frequencies decrease. The incomplete LD between typed and causal SNPs is another inbuilt limitation. It is possible that GWAS hitherto has been underpowered to detect common variants with small effects. Multiple lines of evidence support this contention.

First, international meta-analytical efforts combining ancestral populations have tended to replicate single-population GWAS hits. Marigorta and Navarro (2013) meta-analysed 299 separate GWASs for 28 traits and found consistent correlation of population odds ratios, especially between Europeans and East Asians. Moreover, as cohort numbers have increased, discovered hits have tended to have smaller effect sizes, but not to be less frequent (Marigorta and Navarro, 2013).

Second, common cancer GWAS efforts other than those for CRC, which grew more quickly into international consortia, have seen correspondingly more common variant associations. Al Olama et al. (2014) took the number of prostate cancer-associated loci to 100, explaining 33% of familial risk; Michailidou et al. (2015) increased the number

of breast cancer hits from 79 to 94, which explain 16% of the heritability (although the cause for the disparity in common heritability accessibly explained for different cancers is not as simple as cohort sizes, as discussed in Carvajal Carmona and Tomlinson, 2016).

Third, it's possible to pool all typed common variants to ask what degree of trait heritability they can explain in aggregate, largely thanks to the insight of Yang et al. (2010) that each SNP's effect can be treated as a random, normally distributed nuisance parameter, allowing restricted maximum likelihood estimates of explained phenotypic variance. 295,000 SNPs with MAF$>$ 0.1 explain 45% of the phenotypic variance of height; published GWAS hits alone can account for 5% (Yang et al., 2010). This calculation becomes more difficult for binary traits using the liability scale model of predisposition. The details of this debate are statistically involved to an extent unnecessary for this thesis, and are summarised in figure 7.1. In short, in the special case of a case-control liability score variance, rather than a randomly sampled continuous trait such as height Z-score, two assumptions of restricted maximum likelihood no longer hold. One is the normal distribution of the liability, which balanced case recruitment will destroy (see figure 7.1C, bottom plot). The estimates of common heritability input for bipolar disorder, Crohn's, and type I diabetes by Lee et al. (2011) used a *post hoc* correction for this, but not for its corollary: the skewed liability distribution induces a correlation between genotype and environment (figure 7.1C, with example $r = 0.53$). Golan et al. (2014) show that this subtle error leads to an underestimation of common variant heritability, a bias that increases with sample size. They provided an alternative method, of pairwise regression of genotype correlation on phenotype correlation, to improve common SNP explanations of the heritability of bipolar disorder, Crohn's, type I diabetes, myocardial infarction, multiple sclerosis, and schizophrenia. Common variants explained approximately 60% of the current estimates of heritability for these six traits by Golan et al. (2014)'s method, more than previously thought. Lee et al. (2011)'s method has also been used to estimate how much pooled common variants can explain the heritability of CRC: all SNPs within 250kb of previously reported GWAS hits gave a CRC heritability estimate of 1.14%; all common SNPs (MAF$\leq$ 0.01) together increased this to 7.42%. Two conclusions present themselves. One, a method with a downward

bias nonetheless approaches the lower limit of our existing (and admittedly highly variable, see §1.3.1) CRC heritability estimates. Two, the differences between the Golan et al. (2014) and Lee et al. (2011) methods across multiple traits are considerably less impressive than the difference between any of their estimates and the degree of common heritability explained by GWAS hits, meaning much common variation remains to be associated with disease.

There is a complementary possibility. As well as many additional undiscovered common variant associations, loci already tagged as risk-associated might contribute more to risk than is currently assumed. I alluded to this point in chapter 4. Haplotypic and multi-SNP models enrich accounts of local risk association, even if the constituent variants remain common. Yang et al. (2010) showed that a correction for the incomplete capture of all common SNPs due to imperfect LD between typed and untyped variants could plausibly provide the remainder of the missing heritability for height. The authors stressed that this was merely an exploratory model, to show that common variants and their correlates of similar frequency might form a sufficient model of height heritability. This is a twist on the synthetic association argument for rare variants (Dickson et al., 2010). However, Sampson et al. (2015) used the Lee et al. (2011) method to estimate the common heritability component for 13 different cancers, and these estimates remained static when all SNPs within 250kb of reported susceptibility loci had been removed. I'd maintain, therefore, that although more fine-mapping and haplotypic association modelling will be helpful at GWAS hits, the evidence suggests that a large proportion of missing signal will come from elsewhere.

What of the case for rare variants? Despite the arguments above for untapped common variation, it would be foolish to discount rare genetics. We can now type rare variants in large cohorts, but these collections would have to be enormous to provide sufficient power for individual variant association testing. The only sensible way forward is grouped tests of, for example, all loss of function variants within genes, or, more vaguely, gene sets. The best example I have encountered of this type of study in CRC was performed very early, in 2004, by Fearnhead et al.. The authors manually sequenced five candidate genes (*MLH1*, *MSH2*, *APC*, *CTNNB1*, and *AXIN1*) in 124 patients with

multiple adenomas and 483 random controls, looking for rare coding changes. They reported statistically significant enrichment of *AXIN1* and *MSH2* rare variants in cases. But this is quite different to the design of most recent sequencing papers, and is for a trait that is likely to be more heritable than sporadic CRC. As I've mentioned before, most WGS cohorts are not balanced between cases and controls. Moreover, they rarely pool variants in this way. But this is the sort of study design that is absolutely necessary to supplant or supplement any common GWAS results, as articulated in Zuk et al. (2014). If sporadic liability traits are to be explained by rare variants, we should perform pooled rare variant association studies (RVAS), a subtype of GWAS. Indeed, many GWAS cohorts are already of sufficient size to be well-powered to detect such variants, if they are in gene loss of function groups, provided that selection at those alleles is not too strong (Zuk et al., 2014). But this sort of design is not currently in favour.

Instead, WGS and exome efforts remain wedded to individual variant shortlisting by filtering, either accompanied by linkage study, or approximating linkage with cosegregation analysis. These often occur within patient populations with early-onset cancer or unnaturally many precursor lesions, as in this thesis. This study design reveals their true purpose: to uncover new cancer syndromes or cancer syndrome-like monogenic disorders of moderate penetrance. I have referred to this sort of trait already as quasi-sporadic; Short et al. (2015) call them intermediate phenotypes, and they are no doubt worth discovering. But this means that a more interesting discussion arising from the common/rare variant debate is unfortunately dodged. For in making guesses about the frequency distribution of causative alleles, one is accidentally saying things about the evolutionary forces acting on those alleles, and on the trait in question. It's one thing to suggest that there are additional polyposis syndromes, with perhaps attenuated penetrance or a handful of risk-modifying distant loci. It's another to describe the collective genetics behind a Fisherian complex trait as overwhelmingly rare.

This is an area of inquiry with persistent disagreements, not least over the relationship between selection and allele frequency. Pritchard (2001) argued that the weak selection at moderate effect size variants would actually increase allele frequency, by being sufficient to halt drift to fixation, but incompetent to extinguish the allele against

the ebb of genetic drift. Eyre-Walker (2010) disliked the assumption of Pritchard (2001) that effect on trait and effect on fitness are uncorrelated, and proposed a coupling parameter, τ. τ is an expression of both the degree of causative variant pleiotropy, and the contribution to fitness of the trait caused. If $\tau > 0$ — if the effect of a trait-causing variant is linearly related to selection — then instead most variance is contributed by rare variants. The reasonableness of either common or rare variant models of complex disease might rely, therefore, on how susceptible the disease is to adaptation (or more accurately on τ). As far as this author is aware, the fitness cost of post-reproductive sporadic cancer remains very difficult to measure, and is therefore a matter of some speculation. "The selective disadvantage of [cancer-causing rare] variants will be minimal, especially in relation to the common chronic diseases of the present, which were hardly of any significance until the recent past," say Fearnhead et al. (2004). This contradicts modern followers of Warthin (1913), who predicted "the running-out of a [Lynch syndrome] family line through the gradual development of an inferior stock".

Mancuso et al. (2015) tried to measure τ for prostate cancer causative variants, alighting on a coupling parameter of 0.48. However, this calculation was exclusively in African-ancestry samples, who, having avoided a number of recent population bottlenecks, have a smaller effective population size and more rare variants compared to other human ancestral populations. Their samples of other ancestral populations were insufficiently numerous for a reliable estimate of τ. It's possible that quite different genetic architecture exists for traits in different populations.

The relative contribution of rare and common variants to missing heritability remains moot, although I would lean towards as infinitesimal model of common inputs (Gibson, 2012). Happily, these competing accounts of cancer heredity are testable, by comparison of RVAS to larger common GWAS. But most WGS experiments looking for rare, large effect size variants will have little to say to this debate, and about complex disease genetics more generally. It is for this reason that I say that chapters 5 and 6 (and sequencing) are complementary to chapters 3 and 4 (and GWAS), rather than representative of a methodological improvement. They blur the boundary between Mendelian and sporadic traits, without directly replacing GWAS-provided common inputs. Despite

some implicit promise, current WGS norms do not elaborate on the genetic architecture of sporadic CRC.

A partial explanation for current disfavour with common variation is that it is clinically frustrating. For what can GWAS results currently provide, beyond an excess of unexplained hits, whose biological ascertainment we remain incredibly slow at establishing, and an underwhelming patient stratifier? Even with prostate cancer's tripling of predisposition loci compared to CRC, a prostatic polygenic risk score predicts just a 2.9-fold increase in risk for the top 10% scorers relative to the European population average (Mancuso et al., 2015). Whereas rare variants resemble Mendelian mutations in that they effect a much greater risk and tend to occur more frequently in the somatic setting. This makes them more enticing, especially as pharmacological targets. A standout example of this is the discovery of PPAP. Rare variants in *POLE* explain a heritability syndrome, but also define a subset of colorectal and endometrial cancers with good prognosis (Domingo et al., 2016; Van Gool et al., 2015). This correspondence to somatic mutation is lacking in CRC GWAS results (chapter 6, Machiela et al., 2015).

However, with the discovery of more common variants, the predictive power of polygenic risk scores is likely to improve (Wray et al., 2007). Moreover, the greatest benefit of GWAS is surely the highlighting of new genes or pathways involved in tumourigenesis, as I discussed in §1.4.3, and as allowed the discovery by Gala et al. (2014) of *RNF43* mutation in SPS. Unbiased GWAS will be just as good as sequencing at producing these candidates, if not better because it doesn't skew cases towards more severe phenotypes. It makes little sense to complain that GWAS has provided too few explanations of sporadic cancer when so many reported hits remain completely mysterious. Is rs961253 really a *BMP2* SNP? What is the relationship between the haplotypic signal at *POLD3* and *POLD3* expression? More high-throughput post-GWAS molecular biology is necessary (Stunnenberg and Hubner, 2014). Moreover, if rare variant testing does return to associative case-control cohorts, it will require a considerably improved annotation of the noncoding genome to attempt genome-wide testing. The category pools of variants for RVAS need to be predicted with far more confidence than we can currently muster (Eyre-Walker, 2010; Zuk et al., 2014). The problems of explaining a handful of ncDNA

GWAS results are nothing in comparison. For this reason, post-GWAS should be prioritised. It will complement and enable the eventual exploration of the impact of rare noncoding variation on cancer predisposition.

In summary, the "missing heritability" debate is overblown, but not illegitimate. It's clear that the CDCV model is neither wrong nor exhausted of findings. Common variant trait associations will continue to be discovered. RVAS is also likely to be informative, and a dialogue between CDCV and RVAS GWAS will describe the genetic architecture of sporadic disease liabilities. However, most WGS and exome sequencing papers are not in fact engaged in exploring sporadic disease but are instead implicitly searching for syndromic or quasi-Mendelian pathogenesis, and therefore should not use "missing heritability" to justify themselves, as I was careful to avoid in chapter 6. Moreover, as RVAS becomes feasibly well-powered, category bins of variation will rely on both post-GWAS and ENCODE-like annotative projects. Therefore, WGS neither opposes nor succeeds GWAS. The techniques cooperate.

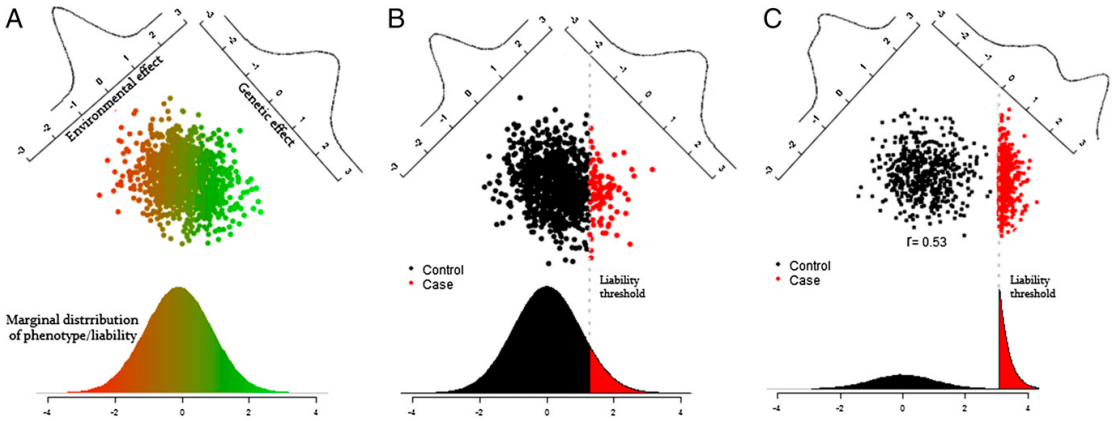


Figure 7.1: Figure from Golan et al. (2014), showing liability distributions in three study designs. *A*, random sample with quantitative phenotype; *B*, random sample with liability phenotype; *C*, case-control study with liability phenotype, causing both a suddenly bimodal liability distribution, as Lee et al. (2011) realised, and also a correlation between genetic and environmental effects, which Lee et al. (2011) do not correct for.

Bibliography

Aaltonen, L. A., Peltomäki, P., Leach, F. S., Sistonen, P., Pylkkänen, L., Mecklin, J. P., Järvinen, H., Powell, S. M., Jen, J., Hamilton, S. R., al. Et, Peltomaki, P., Leach, F. S., Sistonen, P., Pylkkanen, L., Mecklin, J. P., Jarvinen, H., Powell, S. M., Jen, J., and al. Et (1993). Clues to the pathogenesis of familial colorectal cancer. *Science*, 260(5109):812–816.

Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., Hurles, M. E., and McVean, G. A. (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–73.

Abecasis, G. R., Cherny, S. S., Cookson, W. O., and Cardon, L. R. (2002). Merlin–rapid analysis of dense genetic maps using sparse gene flow trees. *Nature genetics*, 30(1):97–101.

Abecasis, G. R. and Wigginton, J. E. (2005). Handling marker-marker linkage disequilibrium: pedigree analysis with clustered markers. *American journal of human genetics*, 77(5):754–767.

Abulí, A., Castells, A., Bujanda, L., Lozano, J. J., Bessa, X., Hernández, C., Álvarez-Urturi, C., Pellisé, M., Esteban-Jurado, C., Hijona, E., Burón, A., Macià, F., Grau, J., Guayta, R., Castellví-Bel, S., and Andreu, M. (2016). Genetic Variants Associated with Colorectal Adenoma Susceptibility. *PloS ONE*, 11(4):e0153084.

Abulí, A., Lozano, J. J., Rodríguez-Soler, M., Jover, R., Bessa, X., Muñoz, J., Esteban-Jurado, C., Fernández-Rozadilla, C., Carracedo, A., Ruiz-Ponte, C., Cubiella, J., esc Balaguer, F., Bujanda, L., Reñé, J. M., Clofent, J., Morillas, J. D., Nicolás-Pérez, D., Xicola, R. M., Llor, X., Piqué, J. M., Andreu, M., Castells, A., and Castellví-Bel, S. (2013). Genetic susceptibility variants associated with colorectal cancer prognosis. *Carcinogenesis*, 34(10):2286–2291.

Adam, R., Spier, I., Zhao, B., Kloth, M., Marquez, J., Hinrichsen, I., Kirfel, J., Tafazzoli, A., Horpaopan, S., Uhlhaas, S., Stienen, D., Friedrichs, N., Altmüller, J., Laner, A., Holzapfel, S., Peters, S., Kayser, K., Thiele, H., Holinski-Feder, E., Marra, G., Kristiansen, G., Nöthen, M. M., Büttner, R., Möslein, G., Betz, R. C., Brieger, A., Lifton, R. P., and Aretz, S. (2016). Exome Sequencing Identifies Biallelic MSH3 Germline Mutations as a Recessive Subtype of Colorectal Adenomatous Polyposis.

Ahlbom, A., Lichtenstein, P., Malmström, H., Feychting, M., Hemminki, K., and Pedersen, N. L. (1997). Cancer in twins: genetic and nongenetic familial risk factors. *Journal of the National Cancer Institute*, 89(4):287–293.

Al Olama, A. A., Kote-Jarai, Z., Berndt, S. I., Conti, D. V., Schumacher, F., Han, Y., Benlloch, S., Hazelett, D. J., Wang, Z., Saunders, E., Leongamornlert, D., Lindstrom, S., Jugurnauth-Little, S., Dadaev, T., Tymrakiewicz, M., Stram, D. O., Rand, K., Wan, P., Stram, A., Sheng, X., Pooler, L. C., Park, K., Xia, L., Tyrer, J., Kolonel, L. N., Le Marchand, L., Hoover, R. N., Machiela, M. J., Yeager, M., Burdette, L., Chung, C. C., Hutchinson, A., Yu, K., Goh, C., Ahmed, M., Govindasami, K., Guy, M., Tammela, T. L. J., Auvinen, A., Wahlfors, T., Schleutker, J., Visakorpi, T., Leinonen, K. A., Xu, J., Aly, M., Donovan, J., Travis, R. C., Key, T. J., Siddiq, A., Canzian, F., Khaw, K.-T., Takahashi, A., Kubo, M., Pharoah, P., Pashayan, N., Weischer, M., Nordestgaard, B. G., Nielsen, S. F., Klarskov, P., Røder, M. A., Iversen, P., Thibodeau, S. N., McDonnell, S. K., Schaid, D. J., Stanford, J. L., Kolb, S., Holt, S., Knudsen, B., Coll, A. H., Gapstur, S. M., Diver, W. R., Stevens, V. L., Maier, C., Luedeke, M., Herkommer, K., Rinckleb, A. E., Strom, S. S., Pettaway, C., Yeboah, E. D., Tettey, Y., Biritwum, R. B., Adjei, A. A., Tay, E., Truelove, A., Niwa, S., Chokkalingam, A. P., Cannon-Albright, L., Cybulski, C., Wokołorczyk, D., Kluźniak, W., Park, J., Sellers, T., Lin, H.-Y., Isaacs, W. B., Partin, A. W., Brenner, H., Dieffenbach, A. K., Stegmaier, C., Chen, C., Giovannucci,

E. L., Ma, J., Stampfer, M., Penney, K. L., Mucci, L., John, E. M., Ingles, S. A., Kittles, R. A., Murphy, A. B., Pandha, H., Michael, A., Kierzek, A. M., Blot, W., Signorello, L. B., Zheng, W., Albanes, D., Virtamo, J., Weinstein, S., Nemesure, B., Carpten, J., Leske, C., Wu, S.-Y., Hennis, A., Kibel, A. S., Rybicki, B. A., Neslund-Dudas, C., Hsing, A. W., Chu, L., Goodman, P. J., Klein, E. A., Zheng, S. L., Batra, J., Clements, J., Spurdle, A., Teixeira, M. R., Paulo, P., Maia, S., Slavov, C., Kaneva, R., Mitev, V., Witte, J. S., Casey, G., Gillanders, E. M., Seminara, D., Riboli, E., Hamdy, F. C., Coetzee, G. A., Li, Q., Freedman, M. L., Hunter, D. J., Muir, K., Gronberg, H., Neal, D. E., Southey, M., Giles, G. G., Severi, G., Cook, M. B., Nakagawa, H., Wiklund, F., Kraft, P., Chanock, S. J., Henderson, B. E., Easton, D. F., Eeles, R. A., Haiman, C. A., Henderson, B. E., Easton, D. F., Eeles, R. A., and Haiman, C. A. (2014). A meta-analysis of 87,040 individuals identifies 23 new susceptibility loci for prostate cancer. *Nature genetics*, 46(10):1103–9.

Al-Tassan, N., Chmiel, N. H., Maynard, J., Fleming, N., Livingston, A. L., Williams, G. T., Hodges, A. K., Davies, D. R., David, S. S., Sampson, J. R., and Cheadle, J. P. (2002). Inherited variants of MYH associated with somatic G:C–>T:A mutations in colorectal tumors. *Nature genetics*, 30(2):227–232.

Al-Tassan, N. A., Whiffin, N., Hosking, F. J., Palles, C., Farrington, S. M., Dobbins, S. E., Harris, R., Gorman, M., Tenesa, A., Meyer, B. F., Wakil, S. M., Kinnersley, B., Campbell, H., Martin, L., Smith, C. G., Idziaszczyk, S., Barclay, E., Maughan, T. S., Kaplan, R., Kerr, R., Kerr, D., Buchanan, D. D., Buchannan, D. D., Ko Win, A., Hopper, J., Jenkins, M., Lindor, N. M., Newcomb, P. A., Gallinger, S., Conti, D., Schumacher, F., Casey, G., Dunlop, M. G., Tomlinson, I. P., Cheadle, J. P., and Houlston, R. S. (2015). A new GWAS and meta-analysis with 1000Genomes imputation identifies novel risk variants for colorectal cancer. *Scientific reports*, 5:10442.

Antoniou, A. C., Pharoah, P. D., Easton, D. F., and Evans, D. G. (2006). BRCA1 and BRCA2 cancer risks. *Journal of Clinical Oncology*, 24(20):3312–3314.

Aust, D. E. and Baretton, G. B. (2010). Serrated polyps of the colon and rectum (hyperplastic polyps, sessile serrated adenomas, traditional serrated adenomas, and mixed polyps)-proposal for diagnostic criteria.

Avsian-Kretchmer, O. and Hsueh, A. J. W. (2004). Comparative genomic analysis of the eight-membered ring cystine knot-containing bone morphogenetic protein antagonists. *Molecular endocrinology (Baltimore, Md.)*, 18(1):1–12.

Baert-Desurmont, S., Charbonnier, F., Houivet, E., Ippolito, L., Mauillon, J., Bougeard, M., Abadie, C., Malka, D., Duffour, J., Desseigne, F., Colas, C., Pujol, P., Lejeune, S., Dugast, C., Buecher, B., Faivre, L., Leroux, D., Gesta, P., Coupier, I., Guimbaud, R., Berthet, P., Manouvrier, S., Cauchin, E., Prieur, F., Laurent-Puig, P., Lebrun, M., Jonveaux, P., Chiesa, J., Caron, O., Morin-Meschin, M.-E., Polycarpe-Osaer, F., Giraud, S., Zaanan, A., Bonnet, D., Mansuy, L., Bonadona, V., El Chehadeh, S., Duhoux, F., Gauthier-Villars, M., Saurin, J.-C., Collonge-Rame, M.-A., Brugières, L., Wang, Q., Bressac-de Paillerets, B., Rey, J.-M., Toulas, C., Buisine, M.-P., Bronner, M., Sokolowska, J., Hardouin, A., Cailleux, A.-F., Sebaoui, H., Blot, J., Tinat, J., Benichou, J., and Frebourg, T. (2015). Clinical relevance of 8q23, 15q13 and 18q21 SNP genotyping to evaluate colorectal cancer risk. *European Journal of Human Genetics*, 24(January):1–7.

Baglioni, S., Melean, G., Gensini, F., Santucci, M., Scatizzi, M., Papi, L., and Genuardi, M. (2005). A kindred with MYH-associated polyposis and pilomatricomas. *American Journal of Medical Genetics*, 134 A(2):212–214.

Bardwell, P. D., Woo, C. J., Wei, K., Li, Z., Martin, A., Sack, S. Z., Parris, T., Edelmann, W., and Scharff, M. D. (2004). Altered somatic hypermutation and reduced class-switch recombination in exonuclease 1-mutant mice. *Nat Immunol*, 5(2):224–229.

Basu, S., Haase, G., and Ben-Ze'ev, A. (2016). Wnt signaling in cancer stem cells and colon cancer metastasis. *F1000Research*, 5(0):699.

Bateson, W., Saunders, E., and Punnett, R. (1905). Experimental studies in the physiology of heredity. *Reports to the Evolution Committee of the Royal Society*, 2:1–55,80–99.

Becker, T., Herold, C., Meesters, C., Mattheisen, M., and Baur, M. P. (2011). Significance levels in genome-wide interaction analysis (GWIA). *Annals of human genetics*, 75(1):29–35.

Belanger, C. F., Hennekens, C. H., Rosner, B., and Speizer, F. E. (1978). The nurses' health study. *The American journal of nursing*, 78(6):1039–40.

Benyamin, B., McRae, A. F., Zhu, G., Gordon, S., Henders, A. K., Palotie, A., Peltonen, L., Martin, N. G., Montgomery, G. W., Whitfield, J. B., and Visscher, P. M. (2009). Variants in TF and HFE Explain approximately 40% of Genetic Variation in Serum-Transferrin Levels. *American Journal of Human Genetics*, 84(1):60–65.

Berger, R., de Vries, S. A. S., Wadman, S. K., Duran, M., Beemer, F. A., de Bree, P. K., Weits-Binnerts, J. J., Penders, T. J., and van der Woude, J. K. (1984). Dihydropyrimidine dehydrogenase deficiency leading to thymine-uraciluria. An inborn error of pyrimidine metabolism. *Clinica Chimica Acta*, 141(2-3):227–234.

Bernstein, A. M., Song, M., Zhang, X., Pan, A., Wang, M., Fuchs, C. S., Le, N., Chan, A. T., Willett, W. C., Ogino, S., Giovannucci, E. L., and Wu, K. (2015). Processed and unprocessed red meat and risk of colorectal cancer: Analysis by tumor location and modification by time. *PLoS ONE*, 10(8):e0135959.

Bertina, R. M., Koeleman, B. P., Koster, T., Rosendaal, F. R., Dirven, R. J., de Ronde, H., van der Velden, P. A., and Reitsma, P. H. (1994). Mutation in blood coagulation factor V associated with resistance to activated protein C. *Nature*, 369(6475):64–7.

Bettington, M., Walker, N., Clouston, A., Brown, I., Leggett, B., and Whitehall, V. (2013). The serrated pathway to colorectal carcinoma: Current concepts and challenges.

Bickeböller, H., Campion, D., Brice, A., Amouyel, P., Hannequin, D., Didierjean, O., Penet, C., Martin, C., Pérez-Tur, J., Michon, A., Dubois, B., Ledoze, F., Thomas-Anterion, C., Pasquier, F., Puel, M., Demonet, J. F., Moreaud, O., Babron, M. C., Meulien, D., Guez, D., Chartier-Harlin, M. C., Frebourg, T., Agid, Y., Martinez, M., and Clerget-Darpoux, F. (1997). Apolipoprotein E and Alzheimer disease: genotype-specific risks by age and sex. *American journal of human genetics*, 60(2):439–446.

Bignold, L. P., Coghlan, B. L. D., and Jersmann, H. P. A. (2006). Hansemann, Boveri, chromosomes and the gametogenesis-related theories of tumours. *Cell Biology International*, 30(7):640–644.

Bodian, D. L., McCutcheon, J. N., Kothiyal, P., Huddleston, K. C., Iyer, R. K., Vockley, J. G., and Niederhuber, J. E. (2014). Germline variation in cancer-susceptibility genes in a healthy, ancestrally diverse cohort: Implications for individual genome sequencing. *PloS ONE*, 9(4):e94554.

Boersma, B. J., Howe, T. M., Goodman, J. E., Yfantis, H. G., Lee, D. H., Chanock, S. J., and Ambs, S. (2006). Association of breast cancer outcome with status of p53 and MDM2 SNP309. *Journal of the National Cancer Institute*, 98(13):911–919.

Bombardier, C., Laine, L., Burgos-Vargas, R., Davis, B., Day, R., Ferraz, M. B., Hawkey, C. J., Hochberg, M. C., Kvien, T. K., Schnitzer, T. J., and Weaver, A. (2006). Response to expression of concern regarding VIGOR study. *The New England journal of medicine*, 354(11):1196–9.

Bombardier, C., Laine, L., Reicin, A., Shapiro, D., Burgos-Vargas, R., Davis, B., Day, R., Ferraz, M. B., Hawkey, C. J., Hochberg, M. C., Kvien, T. K., and Schnitzer, T. J. (2000). Comparison of upper gastrointestinal toxicity of rofecoxib and naproxen in patients with rheumatoid arthritis. VIGOR Study Group. *The New England journal of medicine*, 343(21):1520–8, 2 p following 1528.

Bond, G. L., Hirshfield, K. M., Kirchhoff, T., Alexe, G., Bond, E. E., Robins, H., Bartel, F., Taubert, H., Wuerl, P., Hait, W., Toppmeyer, D., Offit, K., and Levine, A. J. (2006a). MDM2 SNP309 accelerates tumor formation in a gender-specific and hormone-dependent manner. *Cancer Research*, 66(10):5104–5110.

Bond, G. L., Hu, W., Bond, E. E., Robins, H., Lutzker, S. G., Arva, N. C., Bargonetti, J., Bartel, F., Taubert, H., Wuerl, P., Onel, K., Yip, L., Hwang, S. J., Strong, L. C., Lozano, G., and Levine, A. J. (2004). A single nucleotide polymorphism in the MDM2 promoter attenuates the p53 tumor suppressor pathway and accelerates tumor formation in humans. *Cell*, 119(5):591–602.

Bond, G. L., Menin, C., Bertorelle, R., Alhopuro, P., Aaltonen, L. A., and Levine, A. J. (2006b). MDM2 SNP309 accelerates colorectal tumour formation in women. *Journal of medical genetics*, 43(12):950–2.

Boparai, K. S., Mathus-Vliegen, E. M. H., Koornstra, J. J., Nagengast, F. M., van Leerdam, M., van Noesel, C. J. M., Houben, M., Cats, A., van Hest, L. P., Fockens, P., and Dekker, E. (2010). Increased colorectal cancer risk during follow-up in patients with hyperplastic polyposis syndrome: a multicentre cohort study. *Gut*, 59(8):1094–1100.

Botstein, D., White, R. L., Skolnick, M., and Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American journal of human genetics*, 32(3):314–31.

Botteri, E., Iodice, S., Raimondi, S., Maisonneuve, P., and Lowenfels, A. B. (2008). Cigarette Smoking and Adenomatous Polyps: A Meta-analysis. *Gastroenterology*, 134(2):388–395.e3.

Boveri, T. (1904). *Ergebnisse über die Konstitution der chromatischen Substanz des Zellkerns.* Jena, Fischer.

Boveri, T. (1907). Die Entwicklung dispermer Seeigel-Eier. Ein Beitrag zur Befruchtungslehre und zur Theorie des Kerns. *Zellen-Studien, Heft 6.*

Boveri, T. (1914). *Zur Frage der Entstehung maligner Tumoren.* Jena, Gustav Fischer.

Breese, M. R. and Liu, Y. (2013). NGSUtils: a software suite for analyzing and manipulating next-generation sequencing datasets. *Bioinformatics*, 29(4):494–6.

Brenner, H., Chang-claude, J., Seiler, C. M., and Rickert, A. (2011). Protection From Colorectal Cancer After Colonoscopy. *JAMA*, 153(1):22–30.

Bresalier, R. S., Sandler, R. S., Quan, H., Bolognese, J. A., Oxenius, B., Horgan, K., Lines, C., Riddell, R., Morton, D., Lanas, A., Konstam, M. A., and Baron, J. A. (2005). Cardiovascular events associated with rofecoxib in a colorectal adenoma chemoprevention trial. *The New England journal of medicine*, 352(11):1092–102.

Briggs, S. and Tomlinson, I. (2013). Germline and somatic polymerase ϵ and δ mutations define a new class of hypermutated colorectal and endometrial cancers. *The Journal of pathology*, 230(2):148–53.

Broderick, P., Carvajal-Carmona, L., Pittman, A. M., Webb, E., Howarth, K., Rowan, A., Lubbe, S., Spain, S., Sullivan, K., Fielding, S., Jaeger, E., Vijayakrishnan, J., Kemp, Z., Gorman, M., Chandler, I., Papaemmanuil, E., Penegar, S., Wood, W., Sellick, G., Qureshi, M., Teixeira, A., Domingo, E., Barclay, E., Martin, L., Sieber, O., Kerr, D., Gray, R., Peto, J., Cazier, J.-B., Tomlinson, I., and Houlston, R. S. (2007). A genome-wide association study shows that common alleles of SMAD7 influence colorectal cancer risk. *Nature genetics*, 39(11):1315–7.

Bronner, C. E., Baker, S. M., Morrison, P. T., Warren, G., Smith, L. G., Lescoe, M. K., Kane, M., Earabino, C., Lipford, J., Lindblom, A., Tannergård, P., Bollag, R. J., Godwin, A. R., Ward, D. C., Nordenskj Ld, M., Fishel, R., Kolodner, R., and Liskay, R. M. (1994). Mutation in the DNA mismatch repair gene homologue hMLH 1 is associated with hereditary non-polyposis colon cancer. *Nature*, 368(July 2016):258–261.

Budinska, E., Popovici, V., Tejpar, S., D'Ario, G., Lapique, N., Sikora, K. O., Di Narzo, A. F., Yan, P., Graeme Hodgson, J., Weinrich, S., Bosman, F., Roth, A., and Delorenzi, M. (2013). Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer. *Journal of Pathology*, 231(1):63–76.

Burkett, K., Graham, J., and McNeney, B. (2006). hapassoc : Software for Likelihood Inference of Trait Associations with SNP Haplotypes and Other Attributes. *Journal of Statistical Software*, 16(2):1–19.

Burrell, R. A., McClelland, S. E., Endesfelder, D., Groth, P., Weller, M.-C., Shaikh, N., Domingo, E., Kanu, N., Dewhurst, S. M., Gronroos, E., Chew, S. K., Rowan, A. J., Schenk, A., Sheffer, M., Howell, M., Kschischo, M., Behrens, A., Helleday, T., Bartek, J., Tomlinson, I. P., and Swanton, C. (2013). Replication stress links structural and numerical cancer chromosomal instability. *Nature*, 494(7438):492–496.

Butter, F., Davison, L., Viturawong, T., Scheibe, M., Vermeulen, M., Todd, J. A., and Mann, M. (2012). Proteome-wide analysis of disease-associated SNPs that show allele-specific transcription factor binding. *PLoS genetics*, 8(9):e1002982.

Calkins, G. N. (1914). Zur Frage der Entstehung maligner Tumoren. *Science*, 40(1041):1–64.

Carter, H., Marty, R., Hofree, M., Gross, A. M., Jensen, J., Fisch, K. M., Wu, X., Deboever, C., Van Nostrand, E. L., Song, Y., Wheeler, E., Kreisberg, J. F., Lippman, S. M., Yeo, G. W., Gutkind, J. S., and Ideker, T. (2017). Interaction landscape of inherited polymorphisms with somatic events in cancer. *Cancer Discovery*, 7(4):410–423.

Carvajal Carmona, L. G. and Tomlinson, I. (2016). The Hunting of the Snark: Whither Genome-Wide Association Studies for Colorectal Cancer? *Gastroenterology*, 150(7):1528–1530.

Carvajal-Carmona, L. G., Zauber, A. G., Jones, A. M., Howarth, K., Wang, J., Cheng, T., Riddell, R., Lanas, A., Morton, D., Bertagnolli, M. M., and Tomlinson, I. (2013). Much of the genetic risk of colorectal cancer is likely to be mediated through susceptibility to adenomas. *Gastroenterology*, 144(1):53–5.

Castigli, E., Wilson, S. A., Garibyan, L., Rachid, R., Bonilla, F., Schneider, L., and Geha, R. S. (2005). TACI is mutant in common variable immunodeficiency and IgA deficiency. *Nature genetics*, 37(8):829–834.

Castillejo, A., Vargas, G., Castillejo, M. I., Navarro, M., Barberá, V. M., González, S., Hernández-Illán, E., Brunet, J., Ramón Y Cajal, T., Balmaña, J., Oltra, S., Iglesias, S., Velasco, À., Solanes, A., Campos, O., Sánchez Heras, A. B., Gallego, J., Carrasco, E., González Juan, D., Segura, Á., Chirivella, I., Juan, M. J., Tena, I., Lázaro, C., Blanco, I., Pineda, M., Capellá, G., and Soto, J. L. (2014). Prevalence of germline MUTYH mutations among Lynch-like syndrome patients. *European Journal of Cancer*, 50(13):2241–2250.

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, 4(1):7.

Cheah, P. Y., Lim, K. H., Koh, P. K., and Tang, C. L. (2012). The classification of intestinal polyposis. *Nature genetics*, 45(1):2.

Chen, J. and Tian, W. (2016). Explaining the disease phenotype of intergenic SNP through predicted long range regulation. *Nucleic Acids Research*, 44(18):8641–8654.

Chen, J., Xu, J., Zhao, W., Hu, G., Cheng, H., Kang, Y., Xie, Y., and Lu, Y. (2005). Characterization of human LNX, a novel ligand of Numb protein X that is downregulated in human gliomas. *International Journal of Biochemistry and Cell Biology*, 37(11):2273–2283.

Chen, R., Shi, L., Hakenberg, J., Naughton, B., Sklar, P., Zhang, J., Zhou, H., Tian, L., Prakash, O., Lemire, M., Sleiman, P., Cheng, W.-y., Chen, W., Shah, H., Shen, Y., Fromer, M., Omberg, L., Deardorff, M. A., Zackai, E., Bobe, J. R., Levin, E., Hudson, T. J., Groop, L., Wang, J., Hakonarson, H., Wojcicki, A., Diaz, G. A., Edelmann, L., Schadt, E. E., and Friend, S. H. (2016). Analysis of 589,306 genomes identifies individuals resilient to severe Mendelian childhood diseases. *Nature Biotechnology*, 34(5):531–538.

Cheng, T. H., Thompson, D., Painter, J., O'Mara, T., Gorman, M., Martin, L., Palles, C., Jones, A., Buchanan, D. D., Ko Win, A., Hopper, J., Jenkins, M., Lindor, N. M., Newcomb, P. A., Gallinger, S., Conti, D., Schumacher, F., Casey, G., Giles, G. G., Pharoah, P., Peto, J., Cox, A., Swerdlow, A., Couch, F., Cunningham, J. M., Goode, E. L., Winham, S. J., Lambrechts, D., Fasching, P., Burwinkel, B., Brenner, H., Brauch, H., Chang-Claude, J., Salvesen, H. B., Kristensen, V., Darabi, H., Li, J., Liu, T., Lindblom, A., Hall, P., de Polanco, M. E., Sans, M., Carracedo, A., Castellvi-Bel, S., Rojas-Martinez, A., Aguiar Jnr, S., Teixeira, M. R., Dunning, A. M., Dennis, J., Otton, G., Proietto, T., Holliday, E., Attia, J., Ashton, K., Scott, R. J., McEvoy, M., Dowdy, S. C., Fridley, B. L., Werner, H. M., Trovik, J., Njolstad, T. S., Tham, E., Mints, M., Runnebaum, I., Hillemanns, P., Dörk, T., Amant, F., Schrauwen, S., Hein, A., Beckmann, M. W., Ekici, A., Czene, K., Meindl, A., Bolla, M. K., Michailidou, K., Tyrer, J. P., Wang, Q., Ahmed, S., Healey, C. S., Shah, M., Annibali, D., Depreeuw, J., Al-Tassan, N. A., Harris, R., Meyer, B. F., Whiffin, N., Hosking, F. J., Kinnersley, B., Farrington, S. M., Timofeeva, M., Tenesa, A., Campbell, H., Haile, R. W., Hodgson, S., Carvajal-Carmona, L., Cheadle, J. P., Easton, D., Dunlop, M., Houlston, R., Spurdle, A., and Tomlinson, I. (2015). Meta-analysis of genome-wide association studies identifies common susceptibility polymorphisms for colorectal and endometrial cancer near SH2B3 and TSHZ1. *Scientific reports*, 5:17369.

Chih, B., Liu, P., Chinn, Y., Chalouni, C., Komuves, L. G., Hass, P. E., Sandoval, W., and Peterson, A. S. (2011). A ciliopathy complex at the transition zone protects the cilia as a privileged membrane domain. *Nature Cell Biology*, 14(1):61–72.

Christen, W. G., Gaziano, J. M., and Hennekens, C. H. (2000). Design of physicians' health study IIA randomized trial of beta-carotene, vitamins E and C, and multivitamins, in prevention of cancer, cardiovascular disease, and eye disease, and review of results of completed trials. *Annals of Epidemiology*, 10(2):125–134.

Chubb, D., Broderick, P., Dobbins, S. E., Frampton, M., Kinnersley, B., Penegar, S., Price, A., Ma, Y. P., Sherborne, A. L., Palles, C., Timofeeva, M. N., Bishop, D. T., Dunlop, M. G., Tomlinson, I., and Houlston, R. S. (2016). Rare disruptive mutations and their contribution to the heritable risk of colorectal cancer. *Nature communications*, 7:11883.

Corder, E. H., Saunders, A. M., Strittmatter, W. J., Schmechel, D. E., Gaskell, P. C., Small, G. W., Roses, A. D., Haines, J. L., and Pericak-Vance, M. A. (1993). Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science*, 261(5123):921–3.

Correns, C. (1900). G. Mendels Regel Über das Verhalten der Nachkommenschaft der Rassenbastarde. *Berichte der Deutschen Botanischen Gesellschaf*, 18:158–168.

Cortejoso, L., García-González, X., García, M. I., García-Alfonso, P., Sanjurjo, M., and López-Fernández, L. A. (2016). Cost-effectiveness of screening for DPYD polymorphisms to prevent neutropenia in cancer patients treated with fluoropyrimidines. *Pharmacogenomics*, 17(9):979–84.

Cotterchio, M., Manno, M., Klar, N., McLaughlin, J., and Gallinger, S. (2005). Colorectal screening is associated with reduced colorectal cancer risk: A case-control study within the population-based ontario familial colorectal cancer registry. *Cancer Causes and Control*, 16(7):865–875.

Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220.

Cox, D. W., Prat, L., Walshe, J. M., Heathcote, J., and Gaffney, D. (2005). Twenty-four novel mutations in Wilson disease patients of predominantly European ancestry. *Human mutation*, 26(3):280.

Creighton, H. B. and McClintock, B. (1931). A Correlation of Cytological and Genetical Crossing-Over in Zea Mays. *Proceedings of the National Academy of Sciences of the United States of America*, 17(8):492–7.

Crow, J. (1971). Darwinian and Non-Darwinian Evolution Theory. In Le Cam, L., Neyman, J., and Scott, E. L., editors, *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume V*, pages 1–22. University of California Press.

Curfman, G. D., Morrissey, S., and Drazen, J. M. (2005). Expression of concern: Bombardier et al., "Comparison of upper gastrointestinal toxicity of rofecoxib and naproxen in patients with rheumatoid arthritis," N Engl J Med 2000;343:1520-8. *The New England journal of medicine*, 353(26):2813–4.

Cuzick, J., Brentnall, A. R., Segal, C., Byers, H., Reuter, C., Detre, S., Lopez-Knowles, E., Sestak, I., Howell, A., Powles, T. J., Newman, W. G., and Dowsett, M. (2017). Impact of a panel of 88 single nucleotide polymorphisms on the risk of breast cancer in high-risk women: Results from two randomized tamoxifen prevention trials. *Journal of Clinical Oncology*, 35(7):743–750.

Czene, K., Lichtenstein, P., and Hemminki, K. (2002). Environmental and heritable causes of cancer among 9.6 million individuals in the Swedish Family-Cancer Database. *International Journal of Cancer*, 99(2):260–266.

Dai, J., Gu, J., Huang, M., Eng, C., Kopetz, E. S., Ellis, L. M., Hawk, E., and Wu, X. (2012). GWAS-identified colorectal cancer susceptibility loci associated with clinical outcomes. *Carcinogenesis*, 33(7):1327–1331.

Datta, M., Choudhury, A., Lahiri, A., and Bhattacharyya, N. P. (2011). Genome wide gene expression regulation by HIP1 Protein Interactor, HIPPI: prediction and validation. *BMC genomics*, 12(1):463.

D'Auwera, G. (2015). Should I analyze my samples alone or together?

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

de Bakker, P. I. W., Yelensky, R., Pe'er, I., Gabriel, S. B., Daly, M. J., and Altshuler, D. (2005). Efficiency and power in genetic association studies. *Nature genetics*, 37(11):1217–23.

De Sandre-Giovannoli, A., Chaouch, M., Kozlov, S., Vallat, J.-M., Tazir, M., Kassouri, N., Szepetowski, P., Hammadouche, T., Vandenberghe, A., Stewart, C. L., Grid, D., and Lévy, N. (2002). Homozygous defects in LMNA, encoding lamin A/C nuclear-envelope proteins, cause autosomal recessive axonal neuropathy in human (Charcot-Marie-Tooth disorder type 2) and mouse. *American journal of human genetics*, 70(3):726–36.

De Sousa E Melo, F., Wang, X., Jansen, M., Fessler, E., Trinh, A., de Rooij, L. P. M. H., de Jong, J. H., de Boer, O. J., van Leersum, R., Bijlsma, M. F., Rodermond, H., van der Heijden, M., van Noesel, C. J. M., Tuynman, J. B., Dekker, E., Markowetz, F., Medema, J. P., and Vermeulen, L. (2013). Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions. *Nature Medicine*, advance on(5):614–8.

de Voer, R. M., Geurts van Kessel, A., Weren, R. D. A., Ligtenberg, M. J. L., Smeets, D., Fu, L., Vreede, L., Kamping, E. J., Verwiel, E. T. P., Hahn, M.-M., Ariaans, M., Spruijt, L., van Essen, T., Houge, G., Schackert, H. K., Sheng, J. Q., Venselaar, H., van Ravenswaaij-Arts, C. M. A., van Krieken, J. H. J. M., Hoogerbrugge, N., and Kuiper, R. P. (2013). Germline mutations in the spindle assembly checkpoint genes BUB1 and BUB3 are risk factors for colorectal cancer. *Gastroenterology*, 145(3):544–7.

de Voer, R. M., Hahn, M.-M., Mensenkamp, A. R., Hoischen, A., Gilissen, C., Henkes, A., Spruijt, L., van Zelst-Stams, W. A., Kets, C. M., Verwiel, E. T., Nagtegaal, I. D., Schackert, H. K., van Kessel, A. G., Hoogerbrugge, N., Ligtenberg, M. J. L., and Kuiper, R. P. (2015). Deleterious Germline BLM Mutations and the Risk for Early-onset Colorectal Cancer. *Scientific reports*, 5:14060.

de Voer, R. M., Hahn, M.-M., Weren, R. D. A., Mensenkamp, A. R., Gilissen, C., van Zelst-Stams, W. A., Spruijt, L., Kets, C. M., Zhang, J., Venselaar, H., Vreede, L., Schubert, N., Tychon, M., Derks, R., Schackert, H. K., Geurts van Kessel, A., Hoogerbrugge, N., Ligtenberg, M. J. L., and Kuiper, R. P. (2016). Identification of Novel Candidate Genes for Early-Onset Colorectal Cancer Susceptibility. *PLOS Genetics*, 12(2):e1005880.

de Vries, H. G., Collée, J. M., de Walle, H. E., van Veldhuizen, M. H., Smit Sibinga, C. T., Scheffer, H., and ten Kate, L. P. (1997). Prevalence of delta F508 cystic fibrosis carriers in The Netherlands: logistic regression on sex, age, region of residence and number of offspring. *Human genetics*, 99(1):74–9.

Delaneau, O., Marchini, J., and Zagury, J.-F. (2012). A linear complexity phasing method for thousands of genomes. *Nature methods*, 9(2):179–81.

Dempster, E. R. and Lerner, I. M. (1950). Heritability of Threshold Characters. *Genetics*, 35(2):212–236.

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

Dickson, S. P., Wang, K., Krantz, I., Hakonarson, H., and Goldstein, D. B. (2010). Rare variants create synthetic genome-wide associations. *PLoS biology*, 8(1):e1000294.

Dobbins, S. E., Broderick, P., Chubb, D., Kinnersley, B., Sherborne, A. L., and Houlston, R. S. (2016). Undefined familial colorectal cancer and the role of pleiotropism in cancer susceptibility genes. *Familial Cancer*, pages 1–7.

Domingo, E., Freeman-Mills, L., Rayner, E., Glaire, M., Briggs, S., Vermeulen, L., Fessler, E., Medema, J. P., Boot, A., Morreau, H., van Wezel, T., Liefers, G. J., Lothe, R. A., Danielsen, S. A., Sveen, A., Nesbakken, A., Zlobec, I., Lugli, A., Koelzer, V. H., Berger, M. D., Castellví-Bel, S., Muñoz, J., de Bruyn, M., Nijman, H. W., Novelli, M., Lawson, K., Oukrif, D., Frangou, E., Dutton, P., Tejpar, S., Delorenzi, M., Kerr, R., Kerr, D., Tomlinson, I., and Church, D. N. (2016). Somatic POLE

proofreading domain mutation, immune response, and prognosis in colorectal cancer: a retrospective, pooled biomarker study. *The Lancet Gastroenterology and Hepatology*, 1(3):207–216.

Drmanac, R., Sparks, A. B., Callow, M. J., Halpern, A. L., Burns, N. L., Kermani, B. G., Carnevali, P., Nazarenko, I., Nilsen, G. B., Yeung, G., Dahl, F., Fernandez, A., Staker, B., Pant, K. P., Baccash, J., Borcherding, A. P., Brownley, A., Cedeno, R., Chen, L., Chernikoff, D., Cheung, A., Chirita, R., Curson, B., Ebert, J. C., Hacker, C. R., Hartlage, R., Hauser, B., Huang, S., Jiang, Y., Karpinchyk, V., Koenig, M., Kong, C., Landers, T., Le, C., Liu, J., McBride, C. E., Morenzoni, M., Morey, R. E., Mutch, K., Perazich, H., Perry, K., Peters, B. A., Peterson, J., Pethiyagoda, C. L., Pothuraju, K., Richter, C., Rosenbaum, A. M., Roy, S., Shafto, J., Sharanhovich, U., Shannon, K. W., Sheppy, C. G., Sun, M., Thakuria, J. V., Tran, A., Vu, D., Zaranek, A. W., Wu, X., Drmanac, S., Oliphant, A. R., Banyai, W. C., Martin, B., Ballinger, D. G., Church, G. M., and Reid, C. A. (2009). Human Genome Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanoarrays. *Science*, 1469(November):1–7.

Du, M., Jiao, S., Bien, S. A., Gala, M., Abecasis, G. G., Bezieau, S., Brenner, H., Butterbach, K., Caan, B. J., Carlson, C. S. C., Casey, G., Chang-Claude, J., Conti, D. D. V., Curtis, K. R., Duggan, D. D., Gallinger, S., Haile, R. W., Harrison, T. A., Hayes, R. R. B., Hoffmeister, M., Hopper, J. L., Hudson, T. J., Jenkins, M. A., Küry, S., Le Marchand, L., Leal, S. M., Newcomb, P. A., Nickerson, D. A., Potter, J. D., Schoen, R. E., Schumacher, F. R., Seminara, D., Slattery, M. L. M., Hsu, L., Chan, A. T., White, E., Berndt, S. I. S., Peters, U., Lichtenstein, P., Holm, N., Verkasalo, P., Iliadou, A., Kaprio, J., Koskenvuo, M., Dunlop, M., Tenesa, A., Farrington, S., Ballereau, S., Brewster, D., Koessler, T., Peters, U., Jiao, S., Schumacher, F. R., Hutter, C., Aragaki, A., Baron, J., Houlston, R., Cheadle, J., Dobbins, S., Tenesa, A., Jones, A., Howarth, K., Jia, W., Zhang, B., Matsuo, K., Shin, A., Xiang, Y., Jee, S., Dunlop, M., Dobbins, S., Farrington, S., Jones, A., Palles, C., Whiffin, N., Tomlinson, I., Webb, E., Carvajal-Carmona, L., Broderick, P., Howarth, K., Pittman, A., Tomlinson, I., Webb, E., Carvajal-Carmona, L., Broderick, P., Kemp, Z., Spain, S., Zanke, B., Greenwood, C., Rangrej, J., Kustra, R., Tenesa, A., Farrington, S., Haiman, C., Marchand, L. L., Yamamato, J., Stram, D., Sheng, X., Kolonel, L., Hutter, C., Slattery, M. L. M., Duggan, D. D., Muehling, J., Curtin, K., Hsu, L., Kocarnik, J., Hutter, C., Slattery, M. L. M., Berndt, S. I. S., Hsu, L., Duggan, D. D., Tenesa, A., Farrington, S., Prendergast, J., Porteous, M., Walker, M., Haq, N., Tomlinson, I., Carvajal-Carmona, L., Dobbins, S., Tenesa, A., Jones, A., Howarth, K., Houlston, R., Webb, E., Broderick, P., Pittman, A., Bernardo, M. D., Lubbe, S., Jaeger, E., Webb, E., Howarth, K., Carvajal-Carmona, L., Rowan, A., Broderick, P., Broderick, P., Carvajal-Carmona, L., Pittman, A., Webb, E., Howarth, K., Rowan, A., Peters, U., Hutter, C., Hsu, L., Schumacher, F. R., Conti, D. D. V., Carlson, C. S. C., Whiffin, N., Hosking, F., Farrington, S., Palles, C., Dobbins, S., Zgaga, L., Tenesa, A., Dunlop, M., Hirschhorn, J., Daly, M., Freedman, M., Monteiro, A., Gayther, S., Coetzee, G., Risch, A., Plass, C., Lowe, C., Cooper, J., Brusko, T., Walker, N., Smyth, D., Bailey, R., Spain, S., Carvajal-Carmona, L., Howarth, K., Jones, A., Su, Z., Cazier, J., Carvajal-Carmona, L., Cazier, J., Jones, A., Howarth, K., Broderick, P., Pittman, A., Pittman, A., Naranjo, S., Jalava, S., Twiss, P., Ma, Y., Olver, B., Tuupanen, S., Turunen, M., Lehtonen, R., Hallikas, O., Vanharanta, S., Kivioja, T., Pittman, A., Webb, E., Carvajal-Carmona, L., Howarth, K., Bernardo, M. D., Broderick, P., Pittman, A., Naranjo, S., Webb, E., Broderick, P., Lips, E., van Wezel, T., Wang, H., Haiman, C., Burnett, T., Fortini, B., Kolonel, L., Henderson, B., Whiffin, N., Dobbins, S., Hosking, F., Palles, C., Tenesa, A., Wang, Y., Chung, C., Ciampa, J., Yeager, M., Jacobs, K., Berndt, S. I. S., Hayes, R. R. B., Project, C. G., Abecasis, G. G., Auton, A., Brooks, L., DePristo, M., Durbin, R., Hutter, C., Chang-Claude, J., Slattery, M. L. M., Pflugeisen, B., Lin, Y., Duggan, D. D., Price, A., Patterson, N., Plenge, R., Weinblatt, M., Shadick, N., Reich, D., Project, C. G., Abecasis, G. G., Altshuler, D., Auton, A., Brooks, L., Durbin, R., Howie, B., Marchini, J., Stephens, M., Browning, S., Browning, B., Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., Abecasis, G. G., Li, Y., Willer, C., Ding, J., Scheet, P., Abecasis, G. G., Jiao, S., Hsu, L., Hutter, C., Peters, U., Cochran, W., Pruim, R., Welch, R., Sanna, S., Teslovich, T., Chines, P., Gliedt, T., Adzhubei, I., Schmidt, S., Peshkin, L., Ramensky, V., Gerasimova, A., Bork, P., Ward, L., Kellis, M., Ward, L., Kellis, M., Kent, W., Sugnet, C., Furey, T., Roskin, K., Pringle, T., Zahler, A., Consortium, E., Birney, E., Stamatoyannopoulos, J., Dutta, A., Guigo, R., Gingeras, T., Consortium, E., Bernstein, B., Birney, E., Dunham, I., Green, E., Gunter, C., Bernstein, B., Stamatoyannopoulos, J., Costello, J., Ren, B., Milosavljevic, A., Meissner, A., Visel, A., Rubin, E., Pennacchio, L., Visel, A., Blow, M., Li, Z., Zhang, T., Akiyama, J., Holt, A., Siepel, A., Bejerano, G., Pedersen, J., Hinrichs, A., Hou, M., Rosenbloom, K., Pollard, K., Hubisz, M., Rosenbloom, K., Siepel, A., Lappalainen, T., Sammeth, M., Friedlander, M., Hoen, P. t., Monlong, J., Rivas, M., Westra, H., Peters, M., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., Montgomery, S., Sammeth, M., Gutierrez-Arcelus, M., Lach, R., Ingle, C., Nisbett, J., Gibbs, J.,

van der Brug, M., Hernandez, D., Traynor, B., Nalls, M., Lai, S., Schadt, E., Molony, C., Chudin, E., Hao, K., Yang, X., Lum, P., Stranger, B., Nica, A., Forrest, M., Dimas, A., Bird, C., Beazley, C., Reese, M., Eeckman, F., Kulp, D., Haussler, D., Birnbaum, R., Clowney, E., Agamy, O., Kim, M., Zhao, J., Yamanaka, T., Hindorff, L., Sethupathy, P., Junkins, H., Ramos, E., Mehta, J., Coll (2016). Fine-Mapping of Common Genetic Variants Associated with Colorectal Tumor Risk Identified Potential Functional Variants. *PLoS ONE*, 11(7):e0157521.

Dukes, E. (1932). The Classification of Cancer of the Rectum. *The Journal of Pathology and Bacteriology*, 35(3):323–332.

Dunlop, M. G., Dobbins, S. E., Farrington, S. M., Jones, A. M., Palles, C., Whiffin, N., Tenesa, A., Spain, S., Broderick, P., Ooi, L.-Y., Domingo, E., Smillie, C., Henrion, M., Frampton, M., Martin, L., Grimes, G., Gorman, M., Semple, C., Ma, Y. P., Barclay, E., Prendergast, J., Cazier, J.-B., Olver, B., Penegar, S., Lubbe, S., Chander, I., Carvajal-Carmona, L. G., Ballereau, S., Lloyd, A., Vijayakrishnan, J., Zgaga, L., Rudan, I., Theodoratou, E., Starr, J. M., Deary, I., Kirac, I., Kovacević, D., Aaltonen, L. A., Renkonen-Sinisalo, L., Mecklin, J.-P., Matsuda, K., Nakamura, Y., Okada, Y., Gallinger, S., Duggan, D. J., Conti, D., Newcomb, P., Hopper, J., Jenkins, M. A., Schumacher, F., Casey, G., Easton, D., Shah, M., Pharoah, P., Lindblom, A., Liu, T., Smith, C. G., West, H., Cheadle, J. P., Midgley, R., Kerr, D. J., Campbell, H., Tomlinson, I. P., and Houlston, R. S. (2012). Common variation near CDKN1A, POLD3 and SHROOM2 influences colorectal cancer risk. *Nature genetics*, 44(7):770–6.

Dunlop, M. G., Tenesa, A., Farrington, S. M., Ballereau, S., Brewster, D. H., Koessler, T., Pharoah, P., Schafmayer, C., Hampe, J., Völzke, H., Chang-Claude, J., Hoffmeister, M., Brenner, H., von Holst, S., Picelli, S., Lindblom, A., Jenkins, M. A., Hopper, J. L., Casey, G., Duggan, D., Newcomb, P. A., Abulí, A., Bessa, X., Ruiz-Ponte, C., Castellví-Bel, S., Niittymäki, I., Tuupanen, S., Karhu, A., Aaltonen, L., Zanke, B., Hudson, T., Gallinger, S., Barclay, E., Martin, L., Gorman, M., Carvajal-Carmona, L., Walther, A., Kerr, D., Lubbe, S., Broderick, P., Chandler, I., Pittman, A., Penegar, S., Campbell, H., Tomlinson, I., and Houlston, R. S. (2013). Cumulative impact of common genetic variants and other risk factors on colorectal cancer risk in 42 103 individuals. *Gut*, 62(6):871–881.

Durno, C. A., Sherman, P. M., Aronson, M., Malkin, D., Hawkins, C., Bakry, D., Bouffet, E., Gallinger, S., Pollett, A., Campbell, B., and Tabori, U. (2015). Phenotypic and genotypic characterisation of biallelic mismatch repair deficiency (BMMR-D) syndrome. *European Journal of Cancer*, 51(8):977–983.

Eddy, S. R. (2012). The C-value paradox, junk DNA and ENCODE. *Current Biology*, 22(21):R898–R899.

Edelstein, D. L., Axilbund, J. E., Hylind, L. M., Romans, K., Griffin, C. A., Cruz-Correa, M., and Giardiello, F. M. (2013). Serrated polyposis: rapid and relentless development of colorectal neoplasia. *Gut*, 62(3):404–8.

Egger, M., Smith, G. D., Schneider, M., and Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, 315(7109):629–634.

Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H., and Nadeau, J. H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics*, 11(6):446–50.

Elsayed, F. A., Kets, C. M., Ruano, D., van den Akker, B., Mensenkamp, A. R., Schrumpf, M., Nielsen, M., Wijnen, J. T., Tops, C. M., Ligtenberg, M. J., Vasen, H. F., Hes, F. J., Morreau, H., van Wezel, T., Akker, B. V. D., and Arjen, J. (2014). Germline variants in POLE are associated with early onset mismatch repair deficient colorectal cancer. *European journal of human genetics : EJHG*, 23(May):1–5.

Estivill, X., Banccls, C., and Ramos, C. (1997). Geographic distribution and regional origin of 272 cystic fibrosis mutations in European populations. *Human Mutation*, 10(2):135–154.

Eyre-Walker, A. (2010). Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. *Proceedings of the National Academy of Sciences*, 107 Suppl(suppl_1):1752–1756.

Fearnhead, N. S., Wilding, J. L., Winney, B., Tonks, S., Bartlett, S., Bicknell, D. C., Tomlinson, I. P. M., Mortensen, N. J. M., and Bodmer, W. F. (2004). Multiple rare variants in different genes account for multifactorial inherited susceptibility to colorectal adenomas. *Proceedings of the National Academy of Sciences of the United States of America*, 101(45):15992–7.

Feder, J., Gnirke, A., Thomas, W., Tsuchihashi, Z., Ruddy, D., Basava, A., Dormishian, F., Domingo, R., Ellis, M., Fullan, A., Hinton, L., Jones, N., Kimmel, B., Kronmal, G., Lauer, P., Lee, V., Loeb, D., Mapa, F., McClelland, E., Meyer, N., Mintier, G., Moeller, N., Moore, T., Morikang, E., Prass, C., Quintana, L., Starnes, S., Schatzman, R., Brunke, K., Drayna, D., Risch, N., Bacon, B., and Wolff, R. (1996). A novel MHC class Ilike gene is mutated in patients with hereditary haemochromatosis. *Nature Genetics*, 13(4):399–408.

Fehrmann, R. S. N., Jansen, R. C., Veldink, J. H., Westra, H. J., Arends, D., Bonder, M. J., Fu, J., Deelen, P., Groen, H. J. M., Smolonska, A., Weersma, R. K., Hofstra, R. M. W., Buurman, W. A., Rensen, S., Wolfs, M. G. M., Platteel, M., Zhernakova, A., Elbers, C. C., Festen, E. M., Trynka, G., Hofker, M. H., Saris, C. G. J., Ophoff, R. A., van den Berg, L. H., van Heel, D. A., Wijmenga, C., Meerman, G. J., and Franke, L. (2011). Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the hla. *PLoS Genetics*, 7(8):e1002197.

Ferlay, J., Steliarova-Foucher, E., Lortet-Tieulent, J., Rosso, S., Coebergh, J. W. W., Comber, H., Forman, D., and Bray, F. (2013). Cancer incidence and mortality patterns in Europe: estimates for 40 countries in 2012. *European journal of cancer*, 49(6):1374–403.

Figueiredo, J. C., Hsu, L., Hutter, C. M., Lin, Y., Campbell, P. T., Baron, J. A., Berndt, S. I., Jiao, S., Casey, G., Fortini, B., Chan, A. T., Cotterchio, M., Lemire, M., Gallinger, S., Harrison, T. A., Le Marchand, L., Newcomb, P. A., Slattery, M. L., Caan, B. J., Carlson, C. S., Zanke, B. W., Rosse, S. A., Brenner, H., Giovannucci, E. L., Wu, K., Chang-Claude, J., Chanock, S. J., Curtis, K. R., Duggan, D., Gong, J., Haile, R. W., Hayes, R. B., Hoffmeister, M., Hopper, J. L., Jenkins, M. A., Kolonel, L. N., Qu, C., Rudolph, A., Schoen, R. E., Schumacher, F. R., Seminara, D., Stelling, D. L., Thibodeau, S. N., Thornquist, M., Warnick, G. S., Henderson, B. E., Ulrich, C. M., Gauderman, W. J., Potter, J. D., White, E., and Peters, U. (2014). Genome-Wide Diet-Gene Interaction Analyses for Risk of Colorectal Cancer. *PLoS Genetics*, 10(4):e1004228.

Fishel, R., Lescoe, M. K., Rao, M., Copeland, N. G., Jenkins, N. A., Garber, J., Kane, M., and Kolodner, R. (1993). The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer. *Cell*, 75(5):1027–1038.

Fisher, R. (1918). The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Philosophical Transactions of the Royal Society of Edinburgh*, 52:399–433.

Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S. N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E. S., Daly, M. J., and Altshuler, D. (2002). The structure of haplotype blocks in the human genome. *Science*, 296(5576):2225–9.

Gala, M. K., Mizukami, Y., Le, L. P., Moriichi, K., Austin, T., Yamamoto, M., Lauwers, G. Y., Bardeesy, N., and Chung, D. C. (2014). Germline mutations in oncogene-induced senescence pathways are associated with multiple sessile serrated adenomas. *Gastroenterology*, 146(2):520–529.e6.

Gallinger, S., Aronson, M., Shayan, K., Ratcliffe, E. M., Gerstle, J. T., Parkin, P. C., Rothenmund, H., Croitoru, M., Baumann, E., Durie, P. R., Weksberg, R., Pollett, A., Riddell, R. H., Ngan, B. Y., Cutz, E., Lagarde, A. E., and Chan, H. S. L. (2004). Gastrointestinal Cancers and Neurofibromatosis Type 1 Features in Children with a Germline Homozygous MLH1 Mutation. *Gastroenterology*, 126(2):576–585.

Garber, M., Guttman, M., Clamp, M., Zody, M. C., Friedman, N., and Xie, X. (2009). Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics*, 25(12):i54–i62.

Gardiner-Garden, M. and Frommer, M. (1987). CpG Islands in vertebrate genomes. *Journal of Molecular Biology*, 196(2):261–282.

Gatta, G., Capocaccia, R., Coleman, M. P., Gloeckler Ries, L. A., Hakulinen, T., Micheli, A., Sant, M., Verdecchia, A., and Berrino, F. (2000). Toward a comparison of survival in American and European cancer patients. *Cancer*, 89(4):893–900.

Gatta, G., Ciccolallo, L., Capocaccia, R., Coleman, M. P., Hakulinen, T., Møller, H., and Berrino, F. (2003). Differences in colorectal cancer survival between European and US populations: The importance of sub-site and morphology. In *European Journal of Cancer*, volume 39, pages 2214–2222.

Gaudet, M. M., Kuchenbaecker, K. B., Vijai, J., Klein, R. J., Kirchhoff, T., McGuffog, L., Barrowdale, D., Dunning, A. M., Lee, A., Dennis, J., Healey, S., Dicks, E., Soucy, P., Sinilnikova, O. M., Pankratz, V. S., Wang, X., Eldridge, R. C., Tessier, D. C., Vincent, D., Bacot, F., Hogervorst, F. B., Peock, S., Stoppa-Lyonnet, D., Coulet, F., Colas, C., Soubrier, F., Peterlongo, P., Schmutzler, R. K., Nathanson, K. L., Piedmonte, M., Singer, C. F., Thomassen, M., Sokolowska, J., Bronner, M., Hansen, T. V., Neuhausen, S. L., Blanco, I., Greene, M. H., Garber, J., Weitzel, J. N., Andrulis, I. L., Goldgar, D. E., D'Andrea, E., Caldes, T., Nevanlinna, H., Osorio, A., van Rensburg, E. J., Arason, A., Rennert, G., van den Ouweland, A. M., van der Hout, A. H., Kets, C. M., Aalfs, C. M., Wijnen, J. T., Ausems, M. G., Frost, D., Ellis, S., Fineberg, E., Platte, R., Evans, D. G., Jacobs, C., Adlard, J., Tischkowitz, M., Porteous, M. E., Damiola, F., Golmard, L., Barjhoux, L., Longy, M., Belotti, M., Ferrer, S. F., Mazoyer, S., Spurdle, A. B., Manoukian, S., Barile, M., Genuardi, M., Arnold, N., Meindl, A., Sutter, C., Wappenschmidt, B., Domchek, S. M., Pfeiler, G., Friedman, E., Jensen, U. B., Robson, M., Shah, S., Lazaro, C., Mai, P. L., Benitez, J., Southey, M. C., Schmidt, M. K., Fasching, P. A., Peto, J., Humphreys, M. K., Wang, Q., Michailidou, K., Sawyer, E. J., Burwinkel, B., Guénel, P., Bojesen, S. E., Milne, R. L., Brenner, H., Lochmann, M., Brauch, H., Ko, Y. D., Baisch, C., Fischer, H. P., Bruening, T., Pesch, B., Rabstein, S., Spickenheuer, A., Aittomäki, K., Dörk, T., Margolin, S., Mannermaa, A., Lambrechts, D., Chang-Claude, J., Radice, P., Giles, G. G., Haiman, C. A., Winqvist, R., Devillee, P., García-Closas, M., Schoof, N., Hooning, M. J., Cox, A., Pharoah, P. D., Jakubowska, A., Orr, N., González-Neira, A., Pita, G., Alonso, M. R., Hall, P., Couch, F. J., Simard, J., Altshuler, D., Easton, D. F., Chenevix-Trench, G., Antoniou, A. C., Offit, K., Rookus, M. A., Hogervorst, F. B., van Leeuwen, F. E., Verhoef, S., Schmidt, M. K., de Lange, J. L., Collée, J. M., van den Ouweland, A. M., Hooning, M. J., Seynaeve, C., van Deurzen, C. H., van Asperen, C. J., Wijnen, J. T., Tollenaar, R. A., Devilee, P., van Cronenburg, T. C., Kets, C. M., Mensenkamp, A. R., Ausems, M. G., van der Luijt, R. B., Aalfs, C. M., van Os, T. A., Gille, J. J., Waisfisz, Q., Meijers-Heijboer, H. E., Gómez-Garcia, E. B., Blok, M. J., Oosterwijk, J. C., van der Hout, A. H., Mourits, M. J., de Bock, G. H., Vasen, H. F., Peock, S., Frost, D., Ellis, S., Fineberg, E., Platte, R., Miedzybrodzka, Z., Gregory, H., Morrison, P., Jeffers, L., Cole, T., ren Ong, K., Hoffman, J., Donaldson, A., James, M., Tischkowitz, M., Paterson, J., Taylor, A., Murray, A., Rogers, M. T., McCann, E., Kennedy, M. J., Barton, D., Porteous, M., Drummond, S., Brewer, C., Kivuva, E., Searle, A., Goodman, S., Hill, K., Davidson, R., Murday, V., Bradshaw, N., Snadden, L., Longmuir, M., Watt, C., Gibson, S., Haque, E., Tobias, E., Duncan, A., Izatt, L., Jacobs, C., Langman, C., Brady, A., Dorkins, H., Melville, A., Randhawa, K., Barwell, J., Adlard, J., Serra-Feliu, G., Ellis, I., Houghton, C., Lalloo, F., Taylor, J., Side, L., Male, A., Berlin, C., Eason, J., Collier, R., Douglas, F., Claber, O., Jobson, I., Walker, L., McLeod, D., Halliday, D., Durell, S., Stayner, B., Eeles, R. A., Shanley, S., Rahman, N., Houlston, R., Bancroft, E., Page, E., Ardern-Jones, A., Kohut, K., Wiggins, J., Castro, E., Killick, E., Martin, S., Rea, G., Kulkarni, A., Cook, J., Quarrell, O., Bardsley, C., Hodgson, S., Goff, S., Brice, G., Winchester, L., Eddy, C., Tripathi, V., Attard, V., Lehmann, A., Eccles, D., Lucassen, A., Crawford, G., McBride, D., Smalley, S., Sinilnikova, O., Mazoyer, S., Damiola, F., Barjhoux, L., Verny-Pierre, C., Giraud, S., Léone, M., Stoppa-Lyonnet, D., Gauthier-Villars, M., Buecher, B., Houdayer, C., Moncoutier, V., Belotti, M., Tirapo, C., de Pauw, A., Bressac-de Paillerets, B., Caron, O., Bignon, Y. J., Uhrhammer, N., Lasset, C., Bonadona, V., Handallou, S., Hardouin, A., Berthet, P., Sobol, H., Bourdon, V., Noguchi, T., Remenieras, A., Eisinger, F., Coupier, I., Pujol, P., Peyrat, J. P., Fournier, J., Révillion, F., Vennin, P., Adenis, C., Rouleau, E., Lidereau, R., Demange, L., Nogues, C., Muller, D., Fricker, J. P., Barouk-Simonet, E., Bonnet, F., Bubien, V., Sevenet, N., Longy, M., Toulas, C., Guimbaud, R., Gladieff, L., Feillel, V., Dreyfus, H., Rebischung, C., Peysselon, M., Coron, F., Faivre, L., Prieur, F., Lebrun, M., Kientz, C., Ferrer, S. F., Frénay, M., Vénat-Bouvet, L., Delnatte, C., Mortemousque, I., Lynch, H. T., and Snyder, C. L. (2013). Identification of a BRCA2-Specific Modifier Locus at 6p24 Related to Breast Cancer Risk. *PLoS Genetics*, 9(3):e1003173.

Giannakis, M., Hodis, E., Jasmine Mu, X., Yamauchi, M., Rosenbluh, J., Cibulskis, K., Saksena, G., Lawrence, M. S., Qian, Z. R., Nishihara, R., Van Allen, E. M., Hahn, W. C., Gabriel, S. B., Lander,

E. S., Getz, G., Ogino, S., Fuchs, C. S., and Garraway, L. A. (2014). RNF43 is frequently mutated in colorectal and endometrial cancers. *Nature genetics*, 46(12):1264–6.

Gibson, G. (2012). Rare and common variants: twenty arguments. *Nature Reviews Genetics*, 13(2):135–145.

Glatz, K., Pritt, B., Glatz, D., Hartmann, A., O'Brien, M. J., and Blaszyk, H. (2007). A multinational, internet-based assessment of observer variability in the diagnosis of serrated colorectal polyps. *American Journal of Clinical Pathology*, 127(6):938–945.

Glickman, B. W. and Radman, M. (1980). Escherichia coli mutator mutants deficient in methylation-instructed DNA mismatch correction. *Proceedings of the National Academy of Sciences of the United States of America*, 77(2):1063–1067.

Goel, A., Nagasaka, T., Arnold, C. N., Inoue, T., Hamilton, C., Niedzwiecki, D., Compton, C., Mayer, R. J., Goldberg, R., Bertagnolli, M. M., and Boland, C. R. (2007). The CpG Island Methylator Phenotype and Chromosomal Instability Are Inversely Correlated in Sporadic Colorectal Cancer. *Gastroenterology*, 132(1):127–138.

Goellner, E. M., Smith, C. E., Campbell, C. S., Hombauer, H., Desai, A., Putnam, C. D., and Kolodner, R. D. (2014). PCNA and Msh2-Msh6 Activate an Mlh1-Pms1 Endonuclease Pathway Required for Exo1-Independent Mismatch Repair. *Molecular Cell*, 55(2):291–304.

Gohagan, J. K., Prorok, P. C., Hayes, R. B., and Kramer, B.-S. (2000). The Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial of the National Cancer Institute: History, organization, and status. *Controlled Clinical Trials*, 21(6):251S–272S.

Golan, D., Lander, E. S., and Rosset, S. (2014). Measuring missing heritability: Inferring the contribution of common variants. *Proceedings of the National Academy of Sciences*, 111(49):E5272–E5281.

Goldman, H., Ming, S., and Hickock, D. F. (1970). Nature and significance of hyperplastic polyps of the human colon. *Archives of pathology*, 89(4):349–54.

Gong, J., Hutter, C. M., Newcomb, P. A., Ulrich, C. M., Bien, S. A., Campbell, P. T., Baron, J. A., Berndt, S. I., Bezieau, S., Brenner, H., Casey, G., Chan, A. T., Chang-Claude, J., Du, M., Duggan, D., Figueiredo, J. C., Gallinger, S., Giovannucci, E. L., Haile, R. W., Harrison, T. A., Hayes, R. B., Hoffmeister, M., Hopper, J. L., Hudson, T. J., Jeon, J., Jenkins, M. A., Kocarnik, J., Küry, S., Le Marchand, L., Lin, Y., Lindor, N. M., Nishihara, R., Ogino, S., Potter, J. D., Rudolph, A., Schoen, R. E., Schrotz-King, P., Seminara, D., Slattery, M. L., Thibodeau, S. N., Thornquist, M., Toth, R., Wallace, R., White, E., Jiao, S., Lemire, M., Hsu, L., and Peters, U. (2016). Genome-Wide Interaction Analyses between Genetic Variants and Alcohol Consumption and Smoking for Risk of Colorectal Cancer. *PLoS Genetics*, 12(10):e1006296.

Gonzalez-Perez, A. and Lopez-Bigas, N. (2012). Functional impact bias reveals cancer drivers. *Nucleic Acids Research*, 40(21):e169–e169.

Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Tamborero, D., Schroeder, M. P., Jene-Sanz, A., Santos, A., and Lopez-Bigas, N. (2013). IntOGen-mutations identifies cancer drivers across tumor types. *Nature Methods*, 10(11):1081–1082.

Goodman, S. (2014). Polymerase Proofreading-Associated Polyposis (PPAP): an unusual family with the highly penetrant POLE p.Leu424Val mutation with colorectal cancer, polyposis, duodenal, ovarian, gastro oesophageal and ureteric cancers plus diabetes. Technical report.

Gu, L., Frommel, S. C., Oakes, C. C., Simon, R., Grupp, K., Gerig, C. Y., Bar, D., Robinson, M. D., Baer, C., Weiss, M., Gu, Z., Schapira, M., Kuner, R., Sultmann, H., Provenzano, M., Cancer, I. P. o. E. O. P., Yaspo, M.-L., Brors, B., Korbel, J., Schlomm, T., Sauter, G., Eils, R., Plass, C., and Santoro, R. (2015). BAZ2A (TIP5) is involved in epigenetic alterations in prostate cancer and its overexpression predicts disease recurrence. *Nature genetics*, 47(1):22–30.

Gu, Y., Parker, A., Wilson, T. M., Bai, H., Chang, D.-Y., and Lu, A.-L. (2002). Human MutY Homolog, a DNA Glycosylase Involved in Base Excision Repair, Physically and Functionally Interacts with Mismatch Repair Proteins Human MutS Homolog 2/Human MutS Homolog 6. *Journal of Biological Chemistry*, 277(13):11135–11142.

Guinney, J., Dienstmann, R., Wang, X., de Reyniès, A., Schlicker, A., Soneson, C., Marisa, L., Roepman, P., Nyamundanda, G., Angelino, P., Bot, B. M., Morris, J. S., Simon, I. M., Gerster, S., Fessler, E., De Sousa E Melo, F., Missiaglia, E., Ramay, H., Barras, D., Homicsko, K., Maru, D., Manyam, G. C., Broom, B., Boige, V., Perez-Villamil, B., Laderas, T., Salazar, R., Gray, J. W., Hanahan, D., Tabernero, J., Bernards, R., Friend, S. H., Laurent-Puig, P., Medema, J. P., Sadanandam, A., Wessels, L., Delorenzi, M., Kopetz, S., Vermeulen, L., and Tejpar, S. (2015). The consensus molecular subtypes of colorectal cancer. *Nature Medicine*, 21(11):1350–1356.

Gusella, J. F., Wexler, N. S., Conneally, P. M., Naylor, S. L., Anderson, M. A., Tanzi, R. E., Watkins, P. C., Ottina, K., Wallace, M. R., and Sakaguchi, A. Y. (1983). A polymorphic DNA marker genetically linked to Huntington's disease. *Nature*, 306(5940):234–238.

Gylfe, A. E., Kondelin, J., Turunen, M., Ristolainen, H., Katainen, R., Pitkänen, E., Kaasinen, E., Rantanen, V., Tanskanen, T., Varjosalo, M., Lehtonen, H., Palin, K., Taipale, M., Taipale, J., Renkonen-Sinisalo, L., Järvinen, H., Böhm, J., Mecklin, J.-P., Ristimäki, A., Kilpivaara, O., Tuupanen, S., Karhu, A., Vahteristo, P., and Aaltonen, L. A. (2013). Identification of candidate oncogenes in human colorectal cancers with microsatellite instability. *Gastroenterology*, 145(3):540–543.e22.

Hahne, F., Durinck, S., Ivanek, R., Mueller, A., Lianoglou, S., Tan, G., and Parsons, L. (2014). *Gviz: Plotting data and annotation information along genomic coordinates*. R package version 1.12.1.

Halford, S., Sasieni, P., Rowan, A., Wasan, H., Bodmer, W., Talbot, I., Hawkins, N., Ward, R., and Tomlinson, I. (2002). Low-level microsatellite instability occurs in most colorectal cancers and is a nonrandomly distributed quantitative trait. *Cancer Research*, 62(1):53–57.

Hansen, M. F., Johansen, J., Sylvander, A. E., Bjørnevoll, I., Talseth-Palmer, B. A., Lavik, L. A. S., Xavier, A., Engebretsen, L. F., Scott, R. J., Drabløs, F., and Sjursen, W. (2017). Use of multigene-panel identifies pathogenic variants in several CRC-predisposing genes in patients previously tested for Lynch Syndrome. *Clinical Genetics*.

Hao, K., Bossé, Y., Nickle, D. C., Paré, P. D., Postma, D. S., Laviolette, M., Sandford, A., Hackett, T. L., Daley, D., Hogg, J. C., Elliott, W. M., Couture, C., Lamontagne, M., Brandsma, C. A., van den Berge, M., Koppelman, G., Reicin, A. S., Nicholson, D. W., Malkov, V., Derry, J. M., Suver, C., Tsou, J. A., Kulkarni, A., Zhang, C., Vessey, R., Opiteck, G. J., Curtis, S. P., Timens, W., and Sin, D. D. (2012). Lung eQTLs to Help Reveal the Molecular Underpinnings of Asthma. *PLoS Genetics*, 8(11):e1003029.

Hardwick, J. C. H., van Santen, M., van den Brink, G. R., van Deventer, S. J. H., and Peppelenbosch, M. P. (2004). DNA array analysis of the effects of aspirin on colon cancer cells: involvement of Rac1. *Carcinogenesis*, 25(7):1293–8.

Haugen, A. C., Goel, A., Yamada, K., Marra, G., Nguyen, T. P., Nagasaka, T., Kanazawa, S., Koike, J., Kikuchi, Y., Zhong, X., Arita, M., Shibuya, K., Oshimura, M., Hemmi, H., Boland, C. R., and Koi, M. (2008). Genetic instability caused by loss of MutS homologue 3 in human colorectal cancer. *Cancer Research*, 68(20):8465–8472.

Hays, J., Hunt, J. R., Hubbell, F. A., Anderson, G. L., Limacher, M., Allen, C., and Rossouw, J. E. (2003). The Women's Health Initiative recruitment methods and results. *Annals of Epidemiology*, 13(9 SUPPL.):S18–77.

Hazewinkel, Y., Tytgat, K. M., van Eeden, S., Bastiaansen, B., Tanis, P. J., Boparai, K. S., Fockens, P., and Dekker, E. (2014). Incidence of Colonic Neoplasia in Patients With Serrated Polyposis Syndrome Who Undergo Annual Endoscopic Surveillance. *Gastroenterology*, 147(1):88–95.

Hemminki, K., Lönnstedt, I., Vaittinen, P., and Lichtenstein, P. (2001). Estimation of genetic and environmental components in colorectal and lung cancer and melanoma. *Genetic Epidemiology*, 20(1):107–116.

Hershberger, R. E., Parks, S. B., Kushner, J. D., Li, D., Ludwigsen, S., Jakobs, P., Nauman, D., Burgess, D., Partain, J., and Litt, M. (2008). Coding sequence mutations identified in MYH7, TNNT2, SCN5A, CSRP3, LBD3, and TCAP from 313 patients with familial or idiopathic dilated cardiomyopathy. *Clinical and Translational Science*, 1(1):21–26.

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., and Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ (Clinical research ed.)*, 327(7414):557–60.

Higuchi, T., Sugihara, K., and Jass, J. R. (2005). Demographic and pathological characteristics of serrated polyps of colorectum. *Histopathology*, 47(1):32–40.

Hiraki, L. T., Qu, C., Hutter, C. M., Baron, J. A., Berndt, S. I., Bézieau, S., Brenner, H., Caan, B. J., Casey, G., Chang-Claude, J., Chanock, S. J., Conti, D. V., Duggan, D., Fuchs, C. S., Gallinger, S., Giovannucci, E. L., Harrison, T. A., Hayes, R. B., Hazra, A., Henderson, B., Hoffmeister, M., Hopper, J. L., Hudson, T. J., Jenkins, M. A., Küry, S., Marchand, L. L., Lemire, M., Ma, J., Manson, J. A. E., Nan, H., Newcomb, P. A., Ng, K., Potter, J. D., Schoen, R. E., Schumacher, F. R., Seminara, D., Slattery, M. L., Wactawski-Wende, J., White, E., Wu, K., Zanke, B. W., Kraft, P., Peters, U., and Chan, A. T. (2013). Genetic predictors of circulating 25-hydroxyvitamin D and risk of colorectal cancer. *Cancer Epidemiology Biomarkers and Prevention*, 22(11):2037–2046.

Horner, M. J., Ries, L. A. G., Krapcho, M., Neyman, N., Aminou, R., Howlader, N., Altekruse, S. F., Feuer, E. J., Huang, L., and Mariotto, A. (2013). SEER Cancer Statistics Review, 1975-2013. *National Cancer Institute, Bethesda, MD*, 19.

Hoskins, J. M., Ong, P.-S., Keku, T. O., Galanko, J. A., Martin, C. F., Coleman, C. A., Wolfe, M., Sandler, R. S., and McLeod, H. L. (2012). Association of Eleven Common, Low-Penetrance Colorectal Cancer Susceptibility Genetic Variants at Six Risk Loci with Clinical Outcome. *PLoS ONE*, 7(7):e41954.

Houlston, R., Bevan, S., Williams, A., Young, J., Dunlop, M., Rozen, P., Eng, C., Markie, D., Woodford-Richens, K., Rodriguez-Bigas, M. A., Leggett, B., Neale, K., Phillips, R., Sheridan, E., Hodgson, S., Iwama, T., Eccles, D., Bodmer, W., and Tomlinson, I. (1998). Mutations in DPC4 (SMAD4) cause juvenile polyposis syndrome, but only account for a minority of cases. *Human Molecular Genetics*, 7(12):1907–1912.

Houlston, R. S., Cheadle, J., Dobbins, S. E., Tenesa, A., Jones, A. M., Howarth, K., Spain, S. L., Broderick, P., Domingo, E., Farrington, S., Prendergast, J. G. D., Pittman, A. M., Theodoratou, E., Smith, C. G., Olver, B., Walther, A., Barnetson, R. A., Churchman, M., Jaeger, E. E. M., Penegar, S., Barclay, E., Martin, L., Gorman, M., Mager, R., Johnstone, E., Midgley, R., Niittymäki, I., Tuupanen, S., Colley, J., Idziaszczyk, S., Thomas, H. J. W., Lucassen, A. M., Evans, D. G. R., Maher, E. R., Maughan, T., Dimas, A., Dermitzakis, E., Cazier, J.-B., Aaltonen, L. A., Pharoah, P., Kerr, D. J., Carvajal-Carmona, L. G., Campbell, H., Dunlop, M. G., and Tomlinson, I. P. M. (2010). Meta-analysis of three genome-wide association studies identifies susceptibility loci for colorectal cancer at 1q41, 3q26.2, 12q13.13 and 20q13.33. *Nature genetics*, 42(11):973–7.

Houlston, R. S., Webb, E., Broderick, P., Pittman, A. M., Di Bernardo, M. C., Lubbe, S., Chandler, I., Vijayakrishnan, J., Sullivan, K., Penegar, S., Carvajal-Carmona, L., Howarth, K., Jaeger, E., Spain, S. L., Walther, A., Barclay, E., Martin, L., Gorman, M., Domingo, E., Teixeira, A. S., Kerr, D., Cazier, J.-B., Niittymäki, I., Tuupanen, S., Karhu, A., Aaltonen, L. A., Tomlinson, I. P. M., Farrington, S. M., Tenesa, A., Prendergast, J. G. D., Barnetson, R. A., Cetnarskyj, R., Porteous, M. E., Pharoah, P. D. P., Koessler, T., Hampe, J., Buch, S., Schafmayer, C., Tepel, J., Schreiber, S., Völzke, H., Chang-Claude, J., Hoffmeister, M., Brenner, H., Zanke, B. W., Montpetit, A., Hudson, T. J., Gallinger, S., Campbell, H., and Dunlop, M. G. (2008). Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nature genetics*, 40(12):1426–35.

Howe, J. R., Bair, J. L., Sayed, M. G., Anderson, M. E., Mitros, F. A., Petersen, G. M., Velculescu, V. E., Traverso, G., and Vogelstein, B. (2001). Germline mutations of the gene encoding bone morphogenetic protein receptor 1A in juvenile polyposis. *Nature genetics*, 28(2):184–7.

Howe, J. R., Ringold, J. C., Summers, R. W., Mitros, F. A., Nishimura, D., and Stone, E. M. (1998a). A Gene for Familial Juvenile Polyposis Maps to Chromosome 18q21.1. *The American Journal of Human Genetics*, 62(5):1129–1136.

Howe, J. R., Roth, S., Ringold, J. C., Summers, R. W., Järvinen, H. J., Sistonen, P., Tomlinson, I. P., Houlston, R. S., Bevan, S., Mitros, F. A., Stone, E. M., and Aaltonen, L. A. (1998b). Mutations in the SMAD4/DPC4 Gene in Juvenile Polyposis. *Science*, 280(5366):1086–1088.

Hower, V., Evans, S. N., and Pachter, L. (2011). Shape-based peak identification for ChIP-Seq. *BMC bioinformatics*, 12(1):15.

Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., and Abecasis, G. R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature genetics*, 44(8):955–9.

Howie, B. N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*, 5(6):e1000529.

Hsu, D. R., Economides, A. N., Wang, X., Eimon, P. M., and Harland, R. M. (1998). The Xenopus dorsalizing factor Gremlin identifies a novel family of secreted proteins that antagonize BMP activities. *Molecular cell*, 1(5):673–83.

Hsu, L., Jeon, J., Brenner, H., Gruber, S. B., Schoen, R. E., Berndt, S. I., Chan, A. T., Chang-Claude, J., Du, M., Gong, J., Harrison, T. A., Hayes, R. B., Hoffmeister, M., Hutter, C. M., Lin, Y., Nishihara, R., Ogino, S., Prentice, R. L., Schumacher, F. R., Seminara, D., Slattery, M. L., Thomas, D. C., Thornquist, M., Newcomb, P. A., Potter, J. D., Zheng, Y., White, E., and Peters, U. (2015). A model to determine colorectal cancer risk using common genetic susceptibility loci. *Gastroenterology*, 148(7):1330–1339.e14.

Hubner, R. A. and Houlston, R. S. (2007). MTHFR C677T and colorectal cancer risk: A meta-analysis of 25 populations. *International journal of cancer.*, 120(5):1027–35.

Huhn, S., Bevier, M., Pardini, B., Naccarati, A., Vodickova, L., Novotny, J., Vodicka, P., Hemminki, K., and Försti, A. (2014). Colorectal cancer risk and patients' survival: Influence of polymorphisms in genes somatically mutated in colorectal tumors. *Cancer Causes and Control*, 25(6):759–769.

Hunter, D. J., Kraft, P., Jacobs, K. B., Cox, D. G., Yeager, M., Hankinson, S. E., Wacholder, S., Wang, Z., Welch, R., Hutchinson, A., Wang, J., Yu, K., Chatterjee, N., Orr, N., Willett, W. C., Colditz, G. A., Ziegler, R. G., Berg, C. D., Buys, S. S., McCarty, C. A., Feigelson, H. S., Calle, E. E., Thun, M. J., Hayes, R. B., Tucker, M., Gerhard, D. S., Fraumeni, J. F., Hoover, R. N., Thomas, G., and Chanock, S. J. (2007). A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nature genetics*, 39(7):870–4.

Hutter, C. M., Chang-Claude, J., Slattery, M. L., Pflugeisen, B. M., Lin, Y., Duggan, D., Nan, H., Lemire, M., Rangrej, J., Figueiredo, J. C., Jiao, S., Harrison, T. A., Liu, Y., Chen, L. S., Stelling, D. L., Warnick, G. S., Hoffmeister, M., Küry, S., Fuchs, C. S., Giovannucci, E., Hazra, A., Kraft, P., Hunter, D. J., Gallinger, S., Zanke, B. W., Brenner, H., Frank, B., Ma, J., Ulrich, C. M., White, E., Newcomb, P. A., Kooperberg, C., LaCroix, A. Z., Prentice, R. L., Jackson, R. D., Schoen, R. E., Chanock, S. J., Berndt, S. I., Hayes, R. B., Caan, B. J., Potter, J. D., Hsu, L., Bézieau, S., Chan, A. T., Hudson, T. J., and Peters, U. (2012). Characterization of gene-environment interactions for colorectal cancer susceptibility loci. *Cancer Research*, 72(8):2036–2044.

Hutton, S. M. and Spritz, R. A. (2008). A comprehensive genetic study of autosomal recessive ocular albinism in caucasian patients. *Investigative Ophthalmology and Visual Science*, 49(3):868–872.

Hyman, N. H., Anderson, P., and Blasyk, H. (2004). Hyperplastic polyposis and the risk of colorectal cancer. *Diseases of the Colon and Rectum*, 47(12):2101–2104.

Ibáñez-Sanz, G., Díez-Villanueva, A., Alonso, M. H., Rodríguez-Moranta, F., Pérez-Gómez, B., Bustamante, M., Martin, V., Llorca, J., Amiano, P., Ardanaz, E., Tardón, A., Jiménez-Moleón, J. J., Peiró, R., Alguacil, J., Navarro, C., Guinó, E., Binefa, G., Navarro, P. F., Espinosa, A., Dávila-Batista, V., Molina, A. J., Palazuelos, C., Castaño-Vinyals, G., Aragonés, N., Kogevinas, M., Pollán, M., and Moreno, V. (2017). Risk Model for Colorectal Cancer in Spanish Population Using Environmental and Genetic Factors: Results from the MCC-Spain study. *Scientific Reports*, 7(February):43263.

Ikegawa, S. (2012). A short history of the genome-wide association study: where we were and where we are going. *Genomics & informatics*, 10(4):220–5.

Inagaki, N., Hayashi, T., Arimura, T., Koga, Y., Takahashi, M., Shibata, H., Teraoka, K., Chikamori, T., Yamashina, A., and Kimura, A. (2006). αB-crystallin mutation in dilated cardiomyopathy. *Biochemical and Biophysical Research Communications*, 342(2):379–386.

Innocenti, F., Cooper, G. M., Stanaway, I. B., Gamazon, E. R., Smith, J. D., Mirkov, S., Ramirez, J., Liu, W., Lin, Y. S., Moloney, C., Aldred, S. F., Trinklein, N. D., Schuetz, E., Nickerson, D. A., Thummel, K. E., Rieder, M. J., Rettie, A. E., Ratain, M. J., Cox, N. J., and Brown, C. D. (2011). Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue. *PLoS Genetics*, 7(5):e1002078.

Inoue, H., Tanizawa, Y., Wasson, J., Behn, P., Kalidas, K., Bernal-Mizrachi, E., Mueckler, M., Marshall, H., Donis-Keller, H., Crock, P., Rogers, D., Mikuni, M., Kumashiro, H., Higashi, K., Sobue, G., Oka, Y., and Permutt, M. A. (1998). A gene encoding a transmembrane protein is mutated in patients with diabetes mellitus and optic atrophy (Wolfram syndrome). *Nature genetics*, 20(2):143–8.

Ionov, Y., Peinado, M. A., Malkhosyan, S., Shibata, D., and Perucho, M. (1993). Ubiquitous somatic mutations in simple repeated sequences reveal a new mechanism for colonic carcinogenesis. *Nature*, 363(6429):558–561.

Jaeger, E., Leedham, S., Lewis, A., Segditsas, S., Becker, M., Cuadrado, P. R., Davis, H., Kaur, K., Heinimann, K., Howarth, K., East, J., Taylor, J., Thomas, H., and Tomlinson, I. (2012). Hereditary mixed polyposis syndrome is caused by a 40-kb upstream duplication that leads to increased and ectopic expression of the BMP antagonist GREM1. *Nature genetics*, 44(6):699–703.

Jaeger, E., Webb, E., Howarth, K., Carvajal-Carmona, L., Rowan, A., Broderick, P., Walther, A., Spain, S., Pittman, A., Kemp, Z., Sullivan, K., Heinimann, K., Lubbe, S., Domingo, E., Barclay, E., Martin, L., Gorman, M., Chandler, I., Vijayakrishnan, J., Wood, W., Papaemmanuil, E., Penegar, S., Qureshi, M., Farrington, S., Tenesa, A., Cazier, J.-B., Kerr, D., Gray, R., Peto, J., Dunlop, M., Campbell, H., Thomas, H., Houlston, R., and Tomlinson, I. (2008). Common genetic variants at the CRAC1 (HMPS) locus on chromosome 15q13.3 influence colorectal cancer risk. *Nature genetics*, 40(1):26–8.

Jaeger, E., Woodford-Richens, K. L., Lockett, M., Rowan, A. J., Sawyer, E. J., Heinimann, K., Rozen, P., Murday, V. A., Whitelaw, S. C., Ginsberg, A., Atkin, W. S., Lynch, H. T., Southey, M. C., Debinski, H., Eng, C., Bodmer, W. F., Talbot, I. C., Hodgson, S. V., Thomas, H. J. W., and Tomlinson, I. P. M. (2003). An ancestral Ashkenazi haplotype at the HMPS/CRAC1 locus on 15q13-q14 is associated with hereditary mixed polyposis syndrome. *American journal of human genetics*, 72(5):1261–7.

Jamshidi, M., Schmidt, M. K., Dörk, T., Garcia-Closas, M., Heikkinen, T., Cornelissen, S., Van Den Broek, A. J., Schürmann, P., Meyer, A., Park-Simon, T. W., Figueroa, J., Sherman, M., Lissowska, J., Keong, G. T. H., Irwanto, A., Laakso, M., Hautaniemi, S., Aittomäki, K., Blomqvist, C., Liu, J., and Nevanlinna, H. (2013). Germline variation in TP53 regulatory network genes associates with breast cancer survival and treatment outcome. *International Journal of Cancer*, 132(9):2044–2055.

Janssens, F. A. (1909). La Theorie de la Chiasmatypie. *Cellule*, 25:387–411.

Jarvis, D., Mitchell, J. S., Law, P. J., Palin, K., Tuupanen, S., Gylfe, A., Hänninen, U. A., Cajuso, T., Tanskanen, T., Kondelin, J., Kaasinen, E., Sarin, A.-P., Kaprio, J., Eriksson, J. G., Rissanen, H., Knekt, P., Pukkala, E., Jousilahti, P., Salomaa, V., Ripatti, S., Palotie, A., Järvinen, H., Renkonen-Sinisalo, L., Lepistö, A., Böhm, J., Meklin, J.-P., Al-Tassan, N. A., Palles, C., Martin, L., Barclay, E., Farrington, S. M., Timofeeva, M. N., Meyer, B. F., Wakil, S. M., Campbell, H., Smith, C. G., Idziaszczyk, S., Maughan, T. S., Kaplan, R., Kerr, R., Kerr, D., Buchanan, D. D., Win, A. K., Hopper, J. L., Jenkins, M. A., Lindor, N. M., Newcomb, P. A., Gallinger, S., Conti, D., Schumacher, F., Casey, G., Taipale, J., Aaltonen, L. A., Cheadle, J. P., Dunlop, M. G., Tomlinson, I. P., and Houlston, R. S. (2016). Mendelian randomisation analysis strongly implicates adiposity with risk of developing colorectal cancer. *British journal of cancer*.

Jasperson, K. W., Tuohy, T. M., Neklason, D. W., and Burt, R. W. (2010). Hereditary and familial colon cancer. *Gastroenterology*, 138(6):2044–58.

Jass, J. R., Iino, H., Ruszkiewicz, A., Painter, D., Solomon, M. J., Koorey, D. J., Cohn, D., Furlong, K. L., Walsh, M. D., Palazzo, J., Edmonston, T. B., Fishel, R., Young, J., and Leggett, B. A. (2000). Neoplastic progression occurs through mutator pathways in hyperplastic polyposis of the colorectum. *Gut*, 47(1):43–49.

Jeffreys, A., Kauppi, L., and Neumann, R. (2001). Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nature genetics*, 29(2):217–222.

Jemal, A., Siegel, R., Xu, J., and Ward, E. (2011). Cancer statistics, 2010. *CA: a cancer journal for clinicians*, 60(5):277–300.

Jia, W.-H., Zhang, B., Matsuo, K., Shin, A., Xiang, Y.-B., Jee, S. H., Kim, D.-H., Ren, Z., Cai, Q., Long, J., Shi, J., Wen, W., Yang, G., Delahanty, R. J., Ji, B.-T., Pan, Z.-Z., Matsuda, F., Gao, Y.-T., Oh, J. H., Ahn, Y.-O., Park, E. J., Li, H.-L., Park, J. W., Jo, J., Jeong, J.-Y., Hosono, S., Casey, G., Peters, U., Shu, X.-O., Zeng, Y.-X., Zheng, W., Zeng, Y.-X., and Zheng, W. (2012). Genome-wide association analyses in east Asians identify new susceptibility loci for colorectal cancer. *Nature Genetics*, 45(2):191–196.

Johns, L. and Houlston, R. S. (2001). A systematic review and meta-analysis of familial colorectal cancer risk. *The American Journal of Gastroenterology*, 96(10):2992–3003.

Johnson, A. D., Handsaker, R. E., Pulit, S. L., Nizzari, M. M., O'Donnell, C. J., and de Bakker, P. I. W. (2008). SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics*, 24(24):2938–9.

Johnson, R. C., Nelson, G. W., Troyer, J. L., Lautenberger, J. A., Kessing, B. D., Winkler, C. A., and O'Brien, S. J. (2010). Accounting for multiple comparisons in a genome-wide association study (GWAS). *BMC genomics*, 11(1):724.

Johnson, R. E., Klassen, R., Prakash, L., and Prakash, S. (2015). A Major Role of DNA Polymerase d in Replication of Both the Leading and Lagging DNA Strands. *Molecular Cell*, 59(2):163–175.

Jung, K. J., Won, D., Jeon, C., Kim, S., Kim, T. I., Jee, S. H., and Beaty, T. H. (2015). A colorectal cancer prediction model using traditional and genetic risk scores in Koreans. *BMC genetics*, 16(49):1–7.

Kalady, M. F., Jarrar, A., Leach, B., La Guardia, L., O'Malley, M., Eng, C., and Church, J. M. (2011). Defining phenotypes and cancer risk in hyperplastic polyposis syndrome. In *Diseases of the Colon and Rectum*, volume 54, pages 164–170.

Kantor, E. D., Hutter, C. M., Minnier, J., Berndt, S. I., Brenner, H., Caan, B. J., Campbell, P. T., Carlson, C. S., Casey, G., Chan, A. T., Chang-Claude, J., Chanock, S. J., Cotterchio, M., Du, M., Duggan, D., Fuchs, C. S., Giovannucci, E. L., Gong, J., Harrison, T. A., Hayes, R. B., Henderson, B. E., Hoffmeister, M., Hopper, J. L., Jenkins, M. A., Jiao, S., Kolonel, L. N., Marchand, L. L., Lemire, M., Ma, J., Newcomb, P. A., Ochs-Balcom, H. M., Pflugeisen, B. M., Potter, J. D., Rudolph, A., Schoen, R. E., Seminara, D., Slattery, M. L., Stelling, D. L., Thomas, F., Thornquist, M., Ulrich, C. M., Warnick, G. S., Zanke, B. W., Peters, U., Hsu, L., and White, E. (2014). Gene-environment interaction involving recently identified colorectal cancer susceptibility loci. *Cancer Epidemiology Biomarkers and Prevention*, 23(9):1824–1833.

Kaplan, E. L. and Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53(282):457–481.

Kattamuri, C., Luedeke, D. M., Nolan, K., Rankin, S. A., Greis, K. D., Zorn, A. M., and Thompson, T. B. (2012). Members of the DAN family are BMP antagonists that form highly stable noncovalent dimers. *Journal of molecular biology*, 424(5):313–27.

Kerr, D. J., Dunn, J. A., Langman, M. J., Smith, J. L., Midgley, R. S. J., Stanley, A., Stokes, J. C., Julier, P., Iveson, C., Duvvuri, R., and McConkey, C. C. (2007). Rofecoxib and cardiovascular adverse events in adjuvant treatment of colorectal cancer. *The New England journal of medicine*, 357(4):360–9.

Khalid, O., Radaideh, S., Cummings, O. W., O'Brien, M. J., Goldblum, J. R., and Rex, D. K. (2009). Reinterpretation of histology of proximal colon polyps called hyperplastic in 2001. *World journal of gastroenterology*, 15(30):3767–70.

Kheradpour, P. and Kellis, M. (2014). Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Research*, 42(5):2976–2987.

King, M.-C., Marks, J. H., Mandell, J. B., and New York Breast Cancer Study Group (2003). Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2. *Science*, 302(5645):643–6.

Kinnersley, B., Migliorini, G., Broderick, P., Whiffin, N., Dobbins, S. E., Casey, G., Hopper, J., Sieber, O., Lipton, L., Kerr, D. J., Dunlop, M. G., Tomlinson, I. P., and Houlston, R. S. (2012). The TERT variant rs2736100 is associated with colorectal cancer risk. *British journal of cancer*, 107(6):1001–8.

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J.-Y., Sackler, R. S., Haynes, C., Henning, A. K., SanGiovanni, J. P., Mane, S. M., Mayne, S. T., Bracken, M. B., Ferris, F. L., Ott, J., Barnstable, C., and Hoh, J. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science*, 308(5720):385–9.

Knudson, A. G. (1971). Mutation and cancer: statistical study of retinoblastoma. *Proceedings of the National Academy of Sciences of the United States of America*, 68(4):820–3.

Kolonel, L. N., Henderson, B. E., Hankin, J. H., Nomura, A. M., Wilkens, L. R., Pike, M. C., Stram, D. O., Monroe, K. R., Earle, M. E., and Nagamine, F. S. (2000). A multiethnic cohort in Hawaii and Los Angeles: baseline characteristics. *American journal of epidemiology*, 151(4):346–57.

Kong, A. and Cox, N. J. (1997). Allele-sharing models: LOD scores and accurate linkage tests. *American journal of human genetics*, 61(5):1179–88.

Koo, B. K., Spit, M., Jordens, I., Low, T. Y., Stange, D. E., van de Wetering, M., van Es, J. H., Mohammed, S., Heck, A. J., Maurice, M. M., and Clevers, H. (2012). Tumour suppressor RNF43 is a stem-cell E3 ligase that induces endocytosis of Wnt receptors. *Nature*, 488(7413):665–669.

Kosinski, C., Li, V. S. W., Chan, A. S. Y., Zhang, J., Ho, C., Tsui, W. Y., Chan, T. L., Mifflin, R. C., Powell, D. W., Yuen, S. T., Leung, S. Y., and Chen, X. (2007). Gene expression patterns of human colon tops and basal crypts and BMP antagonists as intestinal stem cell niche factors. *Proceedings of the National Academy of Sciences*, 104(39):15418–15423.

Kuchenbaecker, K. B., McGuffog, L., Barrowdale, D., Lee, A., Soucy, P., Dennis, J., Domchek, S. M., Robson, M., Spurdle, A. B., Ramus, S. J., Mavaddat, N., Terry, M. B., Neuhausen, S. L., Schmutzler, R. K., Simard, J., Pharoah, P. D., Offit, K., Couch, F. J., Chenevix-Trench, G., Easton, D. F., and Antoniou, A. C. (2017). Evaluation of Polygenic Risk Scores for Breast and Ovarian Cancer Risk Prediction in BRCA1 and BRCA2 Mutation Carriers. *Journal of the National Cancer Institute*, 109(7).

Kuchenbaecker, K. B., Neuhausen, S. L., Robson, M., Barrowdale, D., McGuffog, L., Mulligan, A. M., Andrulis, I. L., Spurdle, A. B., Schmidt, M. K., Schmutzler, R. K., Engel, C., Wappenschmidt, B., Nevanlinna, H., Thomassen, M., Southey, M., Radice, P., Ramus, S. J., Domchek, S. M., Nathanson, K. L., Lee, A., Healey, S., Nussbaum, R. L., Rebbeck, T. R., Arun, B. K., James, P., Karlan, B. Y., Lester, J., Cass, I., Registry, B. C. F., Terry, M. B., Daly, M. B., Goldgar, D. E., Buys, S. S., Janavicius, R., Tihomirova, L., Tung, N., Dorfling, C. M., van Rensburg, E. J., Steele, L., v O Hansen, T., Ejlertsen, B., Gerdes, A.-M., Nielsen, F. C., Dennis, J., Cunningham, J., Hart, S., Slager, S., Osorio, A., Benitez, J., Duran, M., Weitzel, J. N., Tafur, I., Hander, M., Peterlongo, P., Manoukian, S., Peissel, B., Roversi, G., Scuvera, G., Bonanni, B., Mariani, P., Volorio, S., Dolcetti, R., Varesco, L., Papi, L., Tibiletti, M. G., Giannini, G., Fostira, F., Konstantopoulou, I., Garber, J., Hamann, U., Donaldson, A., Brewer, C., Foo, C., Evans, D. G., Frost, D., Eccles, D., Douglas, F., Brady, A., Cook, J., Tischkowitz, M., Adlard, J., Barwell, J., Ong, K.-r., Walker, L., Izatt, L., Side, L. E., Kennedy, M. J., Rogers, M. T., Porteous, M. E., Morrison, P. J., Platte, R., Eeles, R., Davidson, R., Hodgson, S., Ellis, S., Godwin, A. K., Rhiem, K., Meindl, A., Ditsch, N., Arnold, N., Plendl, H., Niederacher, D., Sutter, C., Steinemann, D., Bogdanova-Markov, N., Kast, K., Varon-Mateeva, R., Wang-Gohrke, S., Gehrig, A., Markiefka, B., Buecher, B., Lefol, C., Stoppa-Lyonnet, D., Rouleau, E., Prieur, F., Damiola, F., Barjhoux, L., Faivre, L., Longy, M., Sevenet, N., Sinilnikova, O. M., Mazoyer, S., Bonadona, V., Caux-Moncoutier, V., Isaacs, C., Van Maerken, T., Claes, K., Piedmonte, M., Andrews, L., Hays, J., Rodriguez, G. C., Caldes, T., de la Hoya, M., Khan, S., Hogervorst, F. B., Aalfs, C. M., de Lange, J., Meijers-Heijboer, H. E., van der Hout, A. H., Wijnen, J. T., van Roozendaal, K., Mensenkamp, A. R., van den Ouweland, A. M., van Deurzen, C. H., van der Luijt, R. B., Olah, E., Diez, O., Lazaro, C., Blanco, I., Teulé, A., Menendez, M., Jakubowska, A., Lubinski, J., Cybulski, C., Gronwald, J., Jaworska-Bieniek, K., Durda, K., Arason, A., Maugard, C., Soucy, P., Montagna, M., Agata, S., Teixeira, M. R., Olswold, C., Lindor, N., Pankratz, V. S., Hallberg, E., Wang, X., Szabo, C. I., Vijai, J., Jacobs, L., Corines, M., Lincoln, A., Berger, A., Fink-Retter, A., Singer, C. F., Rappaport, C., Kaulich, D. G., Pfeiler, G., Tea, M.-K., Phelan, C. M., Mai, P. L., Greene, M. H., Rennert, G., Imyanitov, E. N., Glendon, G., Toland, A. E., Bojesen, A.,

Pedersen, I. S., Jensen, U. B., Caligo, M. A., Friedman, E., Berger, R., Laitman, Y., Rantala, J., Arver, B., Loman, N., Borg, A., Ehrencrona, H., Olopade, O. I., Simard, J., Easton, D. F., Chenevix-Trench, G., Offit, K., Couch, F. J., and Antoniou, A. C. (2014). Associations of common breast cancer susceptibility alleles with risk of breast cancer subtypes in BRCA1 and BRCA2 mutation carriers. *Breast Cancer Research*, 16(6):3416.

Kupfer, S. S., Skol, A. D., Hong, E., Ludvik, A., Kittles, R. A., Keku, T. O., Sandler, R. S., and Ellis, N. A. (2014). Shared and independent colorectal cancer risk alleles in TGFβ-related genes in African and European Americans. *Carcinogenesis*, 35(9):2025–30.

Küry, S., Buecher, B., Robiou-du Pont, S., Scoul, C., Sébille, V., Colman, H., Le Houérou, C., Le Neel, T., Bourdon, J., Faroux, R., Ollivry, J., Lafraise, B., Chupin, L.-D., and Bézieau, S. (2007). Combinations of cytochrome P450 gene polymorphisms enhancing the risk for sporadic colorectal cancer related to red meat consumption. *Cancer epidemiology, biomarkers & prevention*, 16(7):1460–1467.

Lagresle-Peyrou, C., Benjelloun, F., Hue, C., Andre-Schmutz, I., Bonhomme, D., Forveille, M., Beldjord, K., Hacein-Bey-Abina, S., De Villartay, J.-P., Charneau, P., Durandy, A., Fischer, A., and Cavazzana-Calvo, M. (2008). Restoration of human B-cell differentiation into NOD-SCID mice engrafted with gene-corrected CD34+ cells isolated from Artemis or RAG1-deficient patients. *Molecular therapy : the journal of the American Society of Gene Therapy*, 16(2):396–403.

Lander, E. S., Heaford, A., Sheridan, A., Linton, L. M., Birren, B., Subramanian, A., Coulson, A., Nusbaum, C., Zody, M. C., Dunham, A., Baldwin, J., Hunt, A., Devon, K., McMurray, A., Mungall, A., Dewar, K., Doyle, M., Marra, M. A., Fulton, L. a., FitzHugh, W., Chinwalla, A. T., Funke, R., Delehaunty, A., Gage, D., Harris, K., Olsen, A., Gibbs, R. A., Fujiyama, A., Howland, J., Toyoda, A., Kann, L., Rosenthal, A., Lehoczky, J., LeVine, R., Rump, A., McEwan, P., Madan, A., McKernan, K., Federspiel, N. A., Abola, A. P., Meldrim, J., Mesirov, J. P., Evans, G. A., Roe, B. A., Miranda, C., de La Bastide, M., Morris, W., Bailey, J. A., Naylor, J., Bateman, A., Raymond, C., Rosetti, M., Jones, T. A., Kaspryzk, A., Santos, R., McLysaght, A., Sougnez, C., Smit, a. F., Williams, A., Stange-Thomann, N., Stojanovic, N., Felsenfeld, A., Wetterstrand, K. A., Wyman, D., Patrinos, A., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Deadman, R., Deloukas, P., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Jones, M., Lloyd, C., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Mardis, E. R., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Hattori, M., Yada, T., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Platzer, M., Nyakatura, G., Taudien, S., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Qin, S., Davis, R. W., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Shimizu, N., Kawasaki, K., Minoshima, S., Athanasiou, M., Schultz, R., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Kasif, S., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Morgan, M. J., de Jong, P., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y. J., and Szustakowki, J. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

Lane, N. (1976). The Precursor Tissue of Ordinary Large Bowel Cancer. *Cancer Research*, 36(7 PT 2):2669–2672.

Lavoine, N., Colas, C., Muleris, M., Bodo, S., Duval, A., Entz-Werle, N., Coulet, F., Cabaret, O., Andreiuolo, F., Charpy, C., Sebille, G., Wang, Q., Lejeune, S., Buisine, M. P., Leroux, D., Couillault, G., Leverger, G., Fricker, J. P., Guimbaud, R., Mathieu-Dramard, M., Jedraszak, G., Cohen-Hagenauer, O., Guerrini-Rousseau, L., Bourdeaut, F., Grill, J., Caron, O., Baert-Dusermont, S., Tinat, J., Bougeard, G., Frébourg, T., and Brugières, L. (2015). Constitutional mismatch repair deficiency syndrome: clinical description in a French cohort. *Journal of medical genetics*, 52(11):770–8.

Le Marchand, L., Hankin, J. H., Wilkens, L. R., Pierce, L. M., Franke, A., Kolonel, L. N., Seifried, A., Custer, L. J., Chang, W., Lum-Jones, A., and Donlon, T. (2001). Combined effects of well-done red meat, smoking, and rapid N-acetyltransferase 2 and CYP1A2 phenotypes in increasing colorectal cancer risk. *Cancer Epidemiology Biomarkers and Prevention*, 10(12):1259–1266.

Lecarpentier, J., Silvestri, V., Kuchenbaecker, K. B., Barrowdale, D., Dennis, J., McGuffog, L., Soucy, P., Leslie, G., Rizzolo, P., Navazio, A. S., Valentini, V., Zelli, V., Lee, A., Amin Al Olama, A., Tyrer, J. P., Southey, M., John, E. M., Conner, T. A., Goldgar, D. E., Buys, S. S., Janavicius, R., Steele, L., Ding, Y. C., Neuhausen, S. L., Hansen, T. V. O., Osorio, A., Weitzel, J. N., Toss, A., Medici, V., Cortesi, L., Zanna, I., Palli, D., Radice, P., Manoukian, S., Peissel, B., Azzollini, J., Viel, A., Cini, G., Damante, G., Tommasi, S., Peterlongo, P., Fostira, F., Hamann, U., Evans, D. G., Henderson, A., Brewer, C., Eccles, D., Cook, J., Ong, K.-r., Walker, L., Side, L. E., Porteous, M. E., Davidson, R., Hodgson, S., Frost, D., Adlard, J., Izatt, L., Eeles, R., Ellis, S., Tischkowitz, M., EMBRACE, Godwin, A. K., Meindl, A., Gehrig, A., Dworniczak, B., Sutter, C., Engel, C., Niederacher, D., Steinemann, D., Hahnen, E., Hauke, J., Rhiem, K., Kast, K., Arnold, N., Ditsch, N., Wang-Gohrke, S., Wappenschmidt, B., Wand, D., Lasset, C., Stoppa-Lyonnet, D., Belotti, M., Damiola, F., Barjhoux, L., Mazoyer, S., GEMO Study Collaborators, Van Heetvelde, M., Poppe, B., De Leeneer, K., Claes, K. B. M., de la Hoya, M., Garcia-Barberan, V., Caldes, T., Perez Segura, P., Kiiski, J. I., Aittomäki, K., Khan, S., Nevanlinna, H., van Asperen, C. J., HEBON, Vaszko, T., Kasler, M., Olah, E., Balmaña, J., Gutiérrez-Enríquez, S., Diez, O., Teulé, A., Izquierdo, A., Darder, E., Brunet, J., Del Valle, J., Feliubadalo, L., Pujana, M. A., Lazaro, C., Arason, A., Agnarsson, B. A., Johannsson, O. T., Barkardottir, R. B., Alducci, E., Tognazzo, S., Montagna, M., Teixeira, M. R., Pinto, P., Spurdle, A. B., Holland, H., KConFab Investigators, Lee, J. W., Lee, M. H., Lee, J., Kim, S.-W., Kang, E., Kim, Z., Sharma, P., Rebbeck, T. R., Vijai, J., Robson, M., Lincoln, A., Musinsky, J., Gaddam, P., Tan, Y. Y., Berger, A., Singer, C. F., Loud, J. T., Greene, M. H., Mulligan, A. M., Glendon, G., Andrulis, I. L., Toland, A. E., Senter, L., Bojesen, A., Nielsen, H. R., Skytte, A.-B., Sunde, L., Jensen, U. B., Pedersen, I. S., Krogh, L., Kruse, T. A., Caligo, M. A., Yoon, S.-Y., Teo, S.-H., von Wachenfeldt, A., Huo, D., Nielsen, S. M., Olopade, O. I., Nathanson, K. L., Domchek, S. M., Lorenchick, C., Jankowitz, R. C., Campbell, I., James, P., Mitchell, G., Orr, N., Park, S. K., Thomassen, M., Offit, K., Couch, F. J., Simard, J., Easton, D. F., Chenevix-Trench, G., Schmutzler, R. K., Antoniou, A. C., and Ottini, L. (2017). Prediction of Breast and Prostate Cancer Risks in Male BRCA1 and BRCA2 Mutation Carriers Using Polygenic Risk Scores. *Journal of Clinical Oncology*, 35(20):JCO2016694935.

Lee, S. H., Wray, N. R., Goddard, M. E., and Visscher, P. M. (2011). Estimating missing heritability for disease from genome-wide association studies. *American Journal of Human Genetics*, 88(3):294–305.

Leggett, B. A., Devereaux, B., Biden, K., Searle, J., Young, J., and Jass, J. (2001). Hyperplastic polyposis: association with colorectal cancer. *The American journal of surgical pathology*, 25(2):177–184.

Leggett, B. A., Thomas, L. R., Knight, N., Healey, S., Chenevix-Trench, G., and Searle, J. (1993). Exclusion of APC and MCC as the gene defect in one family with familial juvenile polyposis. *Gastroenterology*, 105(5):1313–1316.

Lek, M., Karczewski, K. J., Samocha, K. E., Banks, E., Fennell, T., O, A. H., Ware, J. S., Hill, A. J., Cummings, B. B., Birnbaum, D. P., Kosmicki, J. A., Duncan, L., Zhao, F., Zou, J., Pierce-Hoffman, E., Cooper, D. N., Goldstein, J., Gupta, N., Howrigan, D., Kiezun, A., Stenson, D., Stevens, C., Tiao, G., Tusie-Luna, M. T., Weisburd, B., Wilson, G., Daly, M. J., MacArthur, D. G., Consortium, E. A., Lek, M., Karczewski, K. J., Minikel, E., Samocha, K. E., Banks, E., Fennell, T., O'Donnell-Luria, A., Ware, J. S., Hill, A. J., Cummings, B. B., Tukiainen, T., Birnbaum, D. P., Kosmicki, J. A., Duncan, L., Estrada, K., Zhao, F., Zou, J., Pierce-Hoffman, E., Cooper, D. N., DePristo, M., Do, R., Flannick, J., Fromer, M., Gauthier, L., Goldstein, J., Gupta, N., Howrigan, D., Kiezun, A., Kurki, M., Moonshine, A. L., Natarajan, P., Orozco, L., Peloso, G., Poplin, R., Rivas, M., Ruano-Rubio, V., Ruderfer, D.,

Shakir, K., Stenson, P., Stevens, C., Thomas, B., Tiao, G., Tusie-Luna, M. T., Weisburd, B., Won, H.-H., Yu, D., Altshuler, D., Ardissino, D., Boehnke, M., Danesh, J., Roberto, E., Florez, J., Gabriel, S., Getz, G., Hultman, C., Kathiresan, S., Laakso, M., McCarroll, S., McCarthy, M., McGovern, D., McPherson, R., Neale, B., Palotie, A., Purcell, S., Saleheen, D., Scharf, J., Sklar, P., Patrick, S., Tuomilehto, J., Watkins, H., Wilson, J., Daly, M. J., and MacArthur, D. G. (2015). Analysis of protein-coding genetic variation in 60,706 humans. *bioRxiv*, 536(7616):030338.

Leslie, S., Winney, B., Hellenthal, G., Davison, D., Boumertit, A., Day, T., Hutnik, K., Royrvik, E. C., Cunliffe, B., Wellcome Trust Case Control Consortium 2, International Multiple Sclerosis Genetics Consortium, Lawson, D. J., Falush, D., Freeman, C., Pirinen, M., Myers, S., Robinson, M., Donnelly, P., and Bodmer, W. (2015). The fine-scale genetic structure of the British population. *Nature*, 519(7543):309–314.

Lewis, A., Freeman-Mills, L., de la Calle-Mustienes, E., Giráldez-Pérez, R. M., Davis, H., Jaeger, E., Becker, M., Hubner, N. C., Nguyen, L. N., Zeron-Medina, J., Bond, G., Stunnenberg, H. G., Carvajal, J. J., Gomez-Skarmeta, J. L., Leedham, S., and Tomlinson, I. (2014). A polymorphic enhancer near GREM1 influences bowel cancer risk through differential CDX2 and TCF7L2 binding. *Cell reports*, 8(4):983–90.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–60.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–9.

Li, L., Hamel, N., Baker, K., McGuffin, M. J., Couillard, M., Gologan, A., Marcus, V. A., Chodirker, B., Chudley, A., Stefanovici, C., Durandy, A., Hegele, R. A., Feng, B.-J., Goldgar, D. E., Zhu, J., De Rosa, M., Gruber, S. B., Wimmer, K., Young, B., Chong, G., Tischkowitz, M. D., and Foulkes, W. D. (2015). A homozygous PMS2 founder mutation with an attenuated constitutional mismatch repair deficiency phenotype. *Journal of medical genetics*, 52(5):348–52.

Li, Q., Seo, J.-H., Stranger, B., McKenna, A., Pe'er, I., Laframboise, T., Brown, M., Tyekucheva, S., and Freedman, M. L. (2013). Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. *Cell*, 152(3):633–41.

Lichtenstein, P., Holm, N. V., Verkasalo, P. K., Iliadou, A., Kaprio, J., Koskenvuo, M., Pukkala, E., Skytthe, A., and Hemminki, K. (2000). Environmental and heritable factors in the causation of cancer–analyses of cohorts of twins from Sweden, Denmark, and Finland. *The New England journal of medicine*, 343(2):78–85.

Lindblom, A., Tannergård, P., Werelius, B., and Nordenskjöld, M. (1993). Genetic mapping of a second locus predisposing to hereditary non-polyposis colon cancer. *Nature genetics*, 5(3):279–282.

Liu, J. Z., McRae, A. F., Nyholt, D. R., Medland, S. E., Wray, N. R., Brown, K. M., Hayward, N. K., Montgomery, G. W., Visscher, P. M., Martin, N. G., and Macgregor, S. (2010). A versatile gene-based test for genome-wide association studies. *American journal of human genetics*, 87(1):139–45.

Liu, N. Q., ter Huurne, M., Nguyen, L. N., Peng, T., Wang, S.-Y., Studd, J. B., Joshi, O., Ongen, H., Bramsen, J. B., Yan, J., Andersen, C. L., Taipale, J., Dermitzakis, E. T., Houlston, R. S., Hubner, N. C., and Stunnenberg, H. G. (2017). The non-coding variant rs1800734 enhances DCLK3 expression through long-range interaction and promotes colorectal cancer progression. *Nature Communications*, 8:14418.

Liu, W., Xie, Y., Ma, J., Luo, X., Nie, P., Zuo, Z., Lahrmann, U., Zhao, Q., Zheng, Y., Zhao, Y., Xue, Y., and Ren, J. (2015a). IBS: An illustrator for the presentation and visualization of biological sequences. *Bioinformatics*, 31(20):3359–3361.

Liu, X., Jiang, B., Wang, A., Di, J., Wang, Z., Chen, L., and Su, X. (2015b). GATA2 rs2335052 Polymorphism Predicts the Survival of Patients with Colorectal Cancer. *PloS ONE*, 10(8):e0136020.

Livak, K. J. and Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods (San Diego, Calif.)*, 25(4):402–8.

Longley, M. J., Pierce, A. J., and Modrich, P. (1997). DNA polymerase delta is required for human mismatch repair in vitro. *Journal of Biological Chemistry*, 272(16):10917–10921.

López-García, C., Sansregret, L., Domingo, E., McGranahan, N., Hobor, S., Birkbak, N. J., Horswell, S., Grönroos, E., Favero, F., Rowan, A. J., Matthews, N., Begum, S., Phillimore, B., Burrell, R., Oukrif, D., Spencer-Dene, B., Kovac, M., Stamp, G., Stewart, A., Danielsen, H., Novelli, M., Tomlinson, I., and Swanton, C. (2017). BCL9L Dysfunction Impairs Caspase-2 Expression Permitting Aneuploidy Tolerance in Colorectal Cancer. *Cancer Cell*, 31(1):79–93.

Loudianos, G., Dessi, V., Lovicu, M., Angius, A., Altuntas, B., Giacchino, R., Marazzi, M., Marcellini, M., Sartorelli, M. R., Sturniolo, G. C., Kocak, N., Yuce, A., Akar, N., Pirastu, M., and Cao, A. (1999). Mutation analysis in patients of Mediterranean descent with Wilson disease: identification of 19 novel mutations. *Journal of medical genetics*, 36(11):833–6.

Low, S.-K., Kuchiba, A., Zembutsu, H., Saito, A., Takahashi, A., Kubo, M., Daigo, Y., Kamatani, N., Chiku, S., Totsuka, H., Ohnami, S., Hirose, H., Shimada, K., Okusaka, T., Yoshida, T., Nakamura, Y., and Sakamoto, H. (2010). Genome-wide association study of pancreatic cancer in Japanese population. *PLoS ONE*, 5(7):e11824.

Lucotte, G. (2001). Frequency Analysis and Allele Map in Favor of the Celtic Origin of the C282Y Mutation of Hemochromatosis. *Blood Cells, Molecules, and Diseases*, 27(2):549–556.

Lunter, G. and Goodson, M. (2011). Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome research*, 21(6):936–9.

Lush, J., Lamoreux, W., and Hazel, L. (1948). The heritability of resistance to death in the fowl. *Poult. Sci.*, 27(4):375.

Lynch, H. T. and Krush, A. J. (1971). Cancer family "G" revisited: 1895–1970. *Cancer*, 27(6):1505–1511.

MacArthur, D. G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., Jostins, L., Habegger, L., Pickrell, J. K., Montgomery, S. B., Albers, C. A., Zhang, Z. D., Conrad, D. F., Lunter, G., Zheng, H., Ayub, Q., DePristo, M. A., Banks, E., Hu, M., Handsaker, R. E., Rosenfeld, J. A., Fromer, M., Jin, M., Mu, X. J., Khurana, E., Ye, K., Kay, M., Saunders, G. I., Suner, M.-M. M.-M., Hunt, T., Barnes, I. H. A., Amid, C., Carvalho-Silva, D. R., Bignell, A. H., Snow, C., Yngvadottir, B., Bumpstead, S., Cooper, D. N., Xue, Y., Romero, I. G., Wang, J., Li, Y., Gibbs, R. A., McCarroll, S. A., Dermitzakis, E. T., Pritchard, J. K., Barrett, J. C., Harrow, J., Hurles, M. E., Gerstein, M. B., and Tyler-Smith, C. (2012). A systematic survey of loss-of-function variants in human protein-coding genes. *Science*, 335(6070):823–8.

Machanick, P. and Bailey, T. L. (2011). MEME-ChIP: Motif analysis of large DNA datasets. *Bioinformatics*, 27(12):1696–1697.

Machiela, M. J., Ho, B. M., Fisher, V. A., Hua, X., and Chanock, S. J. (2015). Limited evidence that cancer susceptibility regions are preferential targets for somatic mutation. *Genome biology*, 16:193.

Mancuso, N., Rohland, N., Rand, K. A., Tandon, A., Allen, A., Quinque, D., Mallick, S., Li, H., Stram, A., Sheng, X., Kote-Jarai, Z., Easton, D. F., Eeles, R. A., Le Marchand, L., Lubwama, A., Stram, D., Watya, S., Conti, D. V., Henderson, B., Haiman, C. A., Pasaniuc, B., and Reich, D. (2015). The contribution of rare variation to prostate cancer heritability. *Nature genetics*, advance on.

Mao, B., Wu, W., Davidson, G., Marhold, J., Li, M., Mechler, B. M., Delius, H., Hoppe, D., Stannek, P., Walter, C., Glinka, A., and Niehrs, C. (2002). Kremen proteins are Dickkopf receptors that regulate Wnt/β-catenin signalling. *Nature*, 417(6889):664–667.

Marchini, J. and Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, 11(7):499–511.

Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature genetics*, 39(7):906–13.

Marigorta, U. M. and Navarro, A. (2013). High Trans-ethnic Replicability of GWAS Results Implies Common Causal Variants. *PLoS Genetics*, 9(6):e1003566.

Marisa, L., de Reyniès, A., Duval, A., Selves, J., Gaub, M. P., Vescovo, L., Etienne-Grimaldi, M. C., Schiappa, R., Guenot, D., Ayadi, M., Kirzin, S., Chazal, M., Fléjou, J.-F., Benchimol, D., Berger, A., Lagarde, A., Pencreach, E., Piard, F., Elias, D., Parc, Y., Olschwang, S., Milano, G., Laurent-Puig, P., and Boige, V. (2013). Gene Expression Classification of Colon Cancer into Molecular Subtypes: Characterization, Validation, and Prognostic Value. *PLoS Medicine*, 10(5):e1001453.

Marra, G., Iaccarino, I., Lettieri, T., Roscilli, G., Delmastro, P., and Jiricny, J. (1998). Mismatch repair deficiency associated with overexpression of the MSH3 gene. *Proceedings of the National Academy of Sciences of the United States of America*, 95(15):8568–8573.

Mathieson, I. and McVean, G. (2012). Differential confounding of rare and common variants in spatially structured populations. *Nature genetics*, 44(3):243–6.

Maughan, T. S., Adams, R. A., Smith, C. G., Meade, A. M., Seymour, M. T., Wilson, R. H., Idziaszczyk, S., Harris, R., Fisher, D., Kenny, S. L., Kay, E., Mitchell, J. K., Madi, A., Jasani, B., James, M. D., Bridgewater, J., Kennedy, M. J., Claes, B., Lambrechts, D., Kaplan, R., and Cheadle, J. P. (2011). Addition of cetuximab to oxaliplatin-based first-line combination chemotherapy for treatment of advanced colorectal cancer: results of the randomised phase 3 MRC COIN trial. *Lancet*, 377(9783):2103–14.

Mavaddat, N., Pharoah, P. D. P., Michailidou, K., Tyrer, J., Brook, M. N., Bolla, M. K., Wang, Q., Dennis, J., Dunning, A. M., Shah, M., Luben, R., Brown, J., Bojesen, S. E., Nordestgaard, B. G., Nielsen, S. F., Flyger, H., Czene, K., Darabi, H., Eriksson, M., Peto, J., Dos-Santos-Silva, I., Dudbridge, F., Johnson, N., Schmidt, M. K., Broeks, A., Verhoef, S., Rutgers, E. J., Swerdlow, A., Ashworth, A., Orr, N., Schoemaker, M. J., Figueroa, J., Chanock, S. J., Brinton, L., Lissowska, J., Couch, F. J., Olson, J. E., Vachon, C., Pankratz, V. S., Lambrechts, D., Wildiers, H., Van Ongeval, C., van Limbergen, E., Kristensen, V., Grenaker Alnaes, G., Nord, S., Borresen-Dale, A.-L., Nevanlinna, H., Muranen, T. A., Aittomaki, K., Blomqvist, C., Chang-Claude, J., Rudolph, A., Seibold, P., Flesch-Janys, D., Fasching, P. A., Haeberle, L., Ekici, A. B., Beckmann, M. W., Burwinkel, B., Marme, F., Schneeweiss, A., Sohn, C., Trentham-Dietz, A., Newcomb, P., Titus, L., Egan, K. M., Hunter, D. J., Lindstrom, S., Tamimi, R. M., Kraft, P., Rahman, N., Turnbull, C., Renwick, A., Seal, S., Li, J., Liu, J., Humphreys, K., Benitez, J., Pilar Zamora, M., Arias Perez, J. I., Menendez, P., Jakubowska, A., Lubinski, J., Jaworska-Bieniek, K., Durda, K., Bogdanova, N. V., Antonenkova, N. N., Dork, T., Anton-Culver, H., Neuhausen, S. L., Ziogas, A., Bernstein, L., Devilee, P., Tollenaar, R. A. E. M., Seynaeve, C., van Asperen, C. J., Cox, A., Cross, S. S., Reed, M. W. R., Khusnutdinova, E., Bermisheva, M., Prokofyeva, D., Takhirova, Z., Meindl, A., Schmutzler, R. K., Sutter, C., Yang, R., Schurmann, P., Bremer, M., Christiansen, H., Park-Simon, T.-W., Hillemanns, P., Guenel, P., Truong, T., Menegaux, F., Sanchez, M., Radice, P., Peterlongo, P., Manoukian, S., Pensotti, V., Hopper, J. L., Tsimiklis, H., Apicella, C., Southey, M. C., Brauch, H., Bruning, T., Ko, Y.-D., Sigurdson, A. J., Doody, M. M., Hamann, U., Torres, D., Ulmer, H.-U., Forsti, A., Sawyer, E. J., Tomlinson, I., Kerin, M. J., Miller, N., Andrulis, I. L., Knight, J. A., Glendon, G., Marie Mulligan, A., Chenevix-Trench, G., Balleine, R., Giles, G. G., Milne, R. L., McLean, C., Lindblom, A., Margolin, S., Haiman, C. A., Henderson, B. E., Schumacher, F., Le Marchand, L., Eilber, U., Wang-Gohrke, S., Hooning, M. J., Hollestelle, A., van den Ouweland, A. M. W., Koppert, L. B., Carpenter, J., Clarke, C., Scott, R., Mannermaa, A., Kataja, V., Kosma, V.-M., Hartikainen, J. M., Brenner, H., Arndt, V., Stegmaier, C., Karina Dieffenbach, A., Winqvist, R., Pylkas, K., Jukkola-Vuorinen, A., Grip, M., Offit, K., Vijai, J., Robson, M., Rau-Murthy, R., Dwek, M., Swann, R., Annie Perkins, K., Goldberg, M. S., Labreche, F., Dumont, M., Eccles, D. M., Tapper, W. J., Rafiq, S., John, E. M., Whittemore, A. S., Slager, S., Yannoukakos, D., Toland, A. E., Yao, S., Zheng, W., Halverson, S. L., Gonzalez-Neira, A., Pita, G., Rosario Alonso, M., Alvarez, N., Herrero, D., Tessier, D. C., Vincent, D., Bacot, F., Luccarini, C., Baynes, C., Ahmed, S., Maranian, M., Healey, C. S., Simard, J., Hall, P., Easton, D. F., and Garcia-Closas, M. (2015). Prediction of Breast Cancer Risk Based on Profiling With Common Genetic Variants. *JNCI Journal of the National Cancer Institute*, 107(5):1–15.

McElhinny, S. A. N., Gordenin, D. A., Stith, C. M., Burgers, P. M. J., and Kunkel, T. A. (2008). Division of Labor at the Eukaryotic Replication Fork. *Molecular Cell*, 30:137–144.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, 20(9):1297–303.

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P., Cunningham, F., Eisenstein, M., Weil, M., Chen, A., Visscher, P., Brown, M., McCarthy, M., Yang, J., Pierre, A. S., Génin, E., Zuk, O., Schaffner, S., Samocha, K., Do, R., Hechter, E., Kathiresan, S., Hindorff, L., Sethupathy, P., Junkins, H., Ramos, E., Mehta, J., Collins, F., Puente, X., Beà, S., Valdés-Mas, R., Villamor, N., Gutiérrez-Abril, J., Martín-Subero, J., Gudbjartsson, D., Helgason, H., Gudjonsson, S., Zink, F., Oddson, A., Gylfason, A., Collins, F., Varmus, H., Koepfli, K.-P., Paten, B., O'Brien, S., Cao, J., Schneeberger, K., Ossowski, S., Günther, T., Bender, S., Fitz, J., Daetwyler, H., Capitan, A., Pausch, H., Stothard, P., Binsbergen, R., Brøndum, R., Gonzaga-Jauregui, C., Lupski, J., Gibbs, R., Harrow, J., Frankish, A., Gonzalez, J., Tapanari, E., Diekhans, M., Kokocinski, F., Pruitt, K., Brown, G., Hiatt, S., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., Dalgleish, R., Flicek, P., Cunningham, F., Astashyn, A., Tully, R., Proctor, G., Cunningham, F., Amode, M., Barrell, D., Beal, K., Billis, K., Brent, S., Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P., Cunningham, F., Höglund, J., Sahana, G., Brøndum, R., Guldbrandtsen, B., Buitenhuis, B., Lund, M., Godoy, T., Moreira, G., Boschiero, C., Gheyas, A., Gasparin, G., Paduan, M., Leslie, E., Taub, M., Liu, H., Steinberg, K., Koboldt, D., Zhang, Q., Hou, L., Zhao, H., Saunders, C., Miller, N., Soden, S., Dinwiddie, D., Noll, A., Alnadi, N., Wright, C., Fitzgerald, T., Jones, W., Clayton, S., McRae, J., Kogelenberg, M., McVean, G., Auton, A., Brooks, L., DePristo, M., Durbin, R., Handsaker, R., Paila, U., Chapman, B., Kirchner, R., Quinlan, A., Kersey, P., Allen, J., Christensen, M., Davis, P., Falin, L., Grabmueller, C., Frankish, A., Uszczynska, B., Ritchie, G., Gonzalez, J., Pervouchine, D., Petryszak, R., Rodriguez, J., Maietta, P., Ezkurdia, I., Pietrelli, A., Wesselink, J.-J., Lopez, G., Petryszak, R., Burdett, T., Fiorelli, B., Fonseca, N., Gonzalez-Porta, M., Hastings, E., Kumar, P., Henikoff, S., Ng, P., Adzhubei, I., Schmidt, S., Peshkin, L., Ramensky, V., Gerasimova, A., Bork, P., Gonzalez-Perez, A., Deu-Pons, J., Lopez-Bigas, N., Shihab, H., Gough, J., Cooper, D., Stenson, P., Barker, G., Edwards, K., Schwarz, J., Cooper, D., Schuelke, M., Seelow, D., Ward, L., Kellis, M., Zerbino, D., Wilder, S., Johnson, N., Juettemann, T., Flicek, P., Adams, D., Altucci, L., Antonarakis, S., Ballesteros, J., Beck, S., Bird, A., Romanoski, C., Glass, C., Stunnenberg, H., Wilson, L., Almouzni, G., Cooper, G., Stone, E., Asimenos, G., Green, E., Batzoglou, S., Sidow, A., Ritchie, G., Dunham, I., Zeggini, E., Flicek, P., Kircher, M., Witten, D., Jain, P., O'Roak, B., Cooper, G., Shendure, J., Shihab, H., Gough, J., Mort, M., Cooper, D., Day, I., Gaunt, T., Chen, Y., Cunningham, F., Rios, D., McLaren, W., Smith, J., Pritchard, B., Rios, D., McLaren, W., Chen, Y., Birney, E., Stabenau, A., Flicek, P., Sherry, S., Ward, M., Kholodov, M., Baker, J., Phan, L., Smigielski, E., Forbes, S., Bindal, N., Bamford, S., Cole, C., Kok, C., Beare, D., Lappalainen, I., Lopez, J., Skipper, L., Hefferon, T., Spalding, J., Garner, J., Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Landrum, M., Lee, J., Riley, G., Jang, W., Rubinstein, W., Church, D., Cingolani, P., Platts, A., Wang, L., Coon, M., Nguyen, T., Wang, L., Clarke, L., Zheng-Bradley, X., Smith, R., Kulesha, E., Xiao, C., Toneva, I., Bragin, E., Chatzimichali, E., Wright, C., Hurles, M., Firth, H., Bevan, A., McCarthy, D., Humburg, P., Kanapin, A., Rivas, M., Gaulton, K., Cazier, J.-B., Ardlie, K., Deluca, D., Segrè, A., Sullivan, T., Young, T., Stajich, J., Block, D., Boulez, K., Brenner, S., Chervitz, S., Dagdigian, C., Tan, A., Abecasis, G., Kang, H., Dunnen, J., Antonarakis, S., Wang, K., Li, M., Hakonarson, H., Liu, X., Jian, X., Boerwinkle, E., Jian, X., Boerwinkle, E., and Liu, X. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, 17(1):122.

McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., Wenger, A. M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. *Nature biotechnology*, 28(5):495–501.

Medina-Trillo, C., Aroca-Aguilar, J.-D., Méndez-Hernández, C.-D., Morales, L., García-Antón, M., García-Feijoo, J., and Escribano, J. (2016). Rare FOXC1 variants in congenital glaucoma: identification of translation regulatory sequences. *European journal of human genetics : EJHG*, 24(5):672–80.

Mendel, G. J. (1865). Versuche über Pflanzenhybriden. *Verhandlungen des naturforschenden Vereines in Brünn*, 4:3–47.

Menin, C., Scaini, M. C., De Salvo, G. L., Biscuola, M., Quaggio, M., Esposito, G., Belluco, C., Montagna, M., Agata, S., D'Andrea, E., Nitti, D., Amadori, A., and Bertorelle, R. (2006). Association between MDM2-SNP309 and age at colorectal cancer diagnosis according to p53 mutation status. *Journal of the National Cancer Institute*, 98(4):285–288.

Mensenkamp, A. R., Vogelaar, I. P., Van Zelst-Stams, W. A. G., Goossens, M., Ouchene, H., Hendriks-Cornelissen, S. J. B., Kwint, M. P., Hoogerbrugge, N., Nagtegaal, I. D., and Ligtenberg, M. J. L.

(2014). Somatic mutations in MLH1 and MSH2 are a frequent cause of mismatch-repair deficiency in lynch syndrome-like tumors. *Gastroenterology*, 146(3):643–646.e8.

Merryweather-Clarke, A. T., Pointon, J. J., Shearman, J. D., and Robson, K. J. (1997). Global prevalence of putative haemochromatosis mutations. *Journal of medical genetics*, 34(4):275–278.

Michailidou, K., Beesley, J., Lindstrom, S., Canisius, S., Dennis, J., Lush, M. J., Maranian, M. J., Bolla, M. K., Wang, Q., Shah, M., Perkins, B. J., Czene, K., Eriksson, M., Darabi, H., Brand, J. S., Bojesen, S. E., Nordestgaard, B. G., Flyger, H., Nielsen, S. F., Rahman, N., Turnbull, C., Fletcher, O., Peto, J., Gibson, L., Dos-Santos-Silva, I., Chang-Claude, J., Flesch-Janys, D., Rudolph, A., Eilber, U., Behrens, S., Nevanlinna, H., Muranen, T. A., Aittomäki, K., Blomqvist, C., Khan, S., Aaltonen, K., Ahsan, H., Kibriya, M. G., Whittemore, A. S., John, E. M., Malone, K. E., Gammon, M. D., Santella, R. M., Ursin, G., Makalic, E., Schmidt, D. F., Casey, G., Hunter, D. J., Gapstur, S. M., Gaudet, M. M., Diver, W. R., Haiman, C. A., Schumacher, F., Henderson, B. E., Le Marchand, L., Berg, C. D., Chanock, S. J., Figueroa, J., Hoover, R. N., Lambrechts, D., Neven, P., Wildiers, H., van Limbergen, E., Schmidt, M. K., Broeks, A., Verhoef, S., Cornelissen, S., Couch, F. J., Olson, J. E., Hallberg, E., Vachon, C., Waisfisz, Q., Meijers-Heijboer, H., Adank, M. A., van der Luijt, R. B., Li, J., Liu, J., Humphreys, K., Kang, D., Choi, J.-Y., Park, S. K., Yoo, K.-Y., Matsuo, K., Ito, H., Iwata, H., Tajima, K., Guénel, P., Truong, T., Mulot, C., Sanchez, M., Burwinkel, B., Marme, F., Surowy, H., Sohn, C., Wu, A. H., Tseng, C.-c., Van Den Berg, D., Stram, D. O., González-Neira, A., Benitez, J., Zamora, M. P., Perez, J. I. A., Shu, X.-O., Lu, W., Gao, Y.-T., Cai, H., Cox, A., Cross, S. S., Reed, M. W. R., Andrulis, I. L., Knight, J. A., Glendon, G., Mulligan, A. M., Sawyer, E. J., Tomlinson, I., Kerin, M. J., Miller, N., Lindblom, A., Margolin, S., Teo, S. H., Yip, C. H., Taib, N. A. M., Tan, G.-H., Hooning, M. J., Hollestelle, A., Martens, J. W. M., Collée, J. M., Blot, W., Signorello, L. B., Cai, Q., Hopper, J. L., Southey, M. C., Tsimiklis, H., Apicella, C., Shen, C.-Y., Hsiung, C.-N., Wu, P.-E., Hou, M.-F., Kristensen, V. N., Nord, S., Alnaes, G. I. G., Giles, G. G., Milne, R. L., McLean, C., Canzian, F., Trichopoulos, D., Peeters, P., Lund, E., Sund, M., Khaw, K.-T., Gunter, M. J., Palli, D., Mortensen, L. M., Dossus, L., Huerta, J.-M., Meindl, A., Schmutzler, R. K., Sutter, C., Yang, R., Muir, K., Lophatananon, A., Stewart-Brown, S., Siriwanarangsan, P., Hartman, M., Miao, H., Chia, K. S., Chan, C. W., Fasching, P. A., Hein, A., Beckmann, M. W., Haeberle, L., Brenner, H., Dieffenbach, A. K., Arndt, V., Stegmaier, C., Ashworth, A., Orr, N., Schoemaker, M. J., Swerdlow, A. J., Brinton, L., Garcia-Closas, M., Zheng, W., Halverson, S. L., Shrubsole, M., Long, J., Goldberg, M. S., Labrèche, F., Dumont, M., Winqvist, R., Pylkäs, K., Jukkola-Vuorinen, A., Grip, M., Brauch, H., Hamann, U., Brüning, T., Radice, P., Peterlongo, P., Manoukian, S., Bernard, L., Bogdanova, N. V., Dörk, T., Mannermaa, A., Kataja, V., Kosma, V.-M., Hartikainen, J. M., Devilee, P., Tollenaar, R. A. E. M., Seynaeve, C., Van Asperen, C. J., Jakubowska, A., Lubinski, J., Jaworska, K., Huzarski, T., Sangrajrang, S., Gaborieau, V., Brennan, P., McKay, J., Slager, S., Toland, A. E., Ambrosone, C. B., Yannoukakos, D., Kabisch, M., Torres, D., Neuhausen, S. L., Anton-Culver, H., Luccarini, C., Baynes, C., Ahmed, S., Healey, C. S., Tessier, D. C., Vincent, D., Bacot, F., Pita, G., Alonso, M. R., Álvarez, N., Herrero, D., Simard, J., Pharoah, P. P. D. P., Kraft, P., Dunning, A. M., Chenevix-Trench, G., Hall, P., and Easton, D. F. (2015). Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. *Nature Genetics*, 47(4):373–80.

Midgley, R. S., McConkey, C. C., Johnstone, E. C., Dunn, J. A., Smith, J. L., Grumett, S. A., Julier, P., Iveson, C., Yanagisawa, Y., Warren, B., Langman, M. J., and Kerr, D. J. (2010). Phase III randomized trial assessing rofecoxib in the adjuvant setting of colorectal cancer: final results of the VICTOR trial. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 28(30):4575–80.

Miyaki, M., Konishi, M., Tanaka, K., Kikuchi-Yanoshita, R., Muraoka, M., Yasuno, M., Igari, T., Koike, M., Chiba, M., and Mori, T. (1997). Germline mutation of MSH6 as the cause of hereditary nonpolyposis colorectal cancer. *Nature Genetics*, 17(3):271–272.

Morgan, T. H. (1910a). Chromosomes and Heredity. *The American Naturalist*, 44(524):449–496.

Morgan, T. H. (1910b). Sex-limited inheritance in Drosophila. *Science*, 32:120–122.

Morgan, T. H. (1911). Random Segregation Versus Coupling In Mendelian Inheritance. *Science*, 34(873):384.

Morris, E. J. A., Penegar, S., Whiffin, N., Broderick, P., Bishop, D. T., Northwood, E., Quirke, P., Finan, P., and Houlston, R. S. (2015). A retrospective observational study of the relationship between single nucleotide polymorphisms associated with the risk of developing Colorectal cancer and survival. *PLoS ONE*, 10(2):e0117816.

Müller, J. S., Baumeister, S. K., Rasic, V. M., Krause, S., Todorovic, S., Kugler, K., Müller-Felber, W., Abicht, A., and Lochmüller, H. (2006). Impaired receptor clustering in congenital myasthenic syndrome with novel RAPSN mutations. *Neurology*, 67(7):1159–1164.

Muzny, D. D. M., Bainbridge MN, Chang K, Dinh HH, Drummond JA, Fowler G, Kovar CL, Lewis LR, Morgan MB, Newsham IF, Reid JG, Santibanez J, Shinbrot E, Trevino LR, Wu YQ, Wang M, Gunaratne P, Donehower LA, Creighton CJ, Wheeler DA, Gibbs RA, Lawrence MS, Voet D, Jing R, Ci, P., Bainbridge, M. N., Chang, K., Dinh, H. H., Drummond, J. A., Fowler, G., Kovar, C. L., Lewis, L. R., Morgan, M. B., Newsham, I. F., Reid, J. G., Santibanez, J., Shinbrot, E., Trevino, L. R., Wu, Y.-Q., Wang, M., Gunaratne, P., Donehower, L. A., Creighton, C. J., Wheeler, D. A., Gibbs, R. A., Lawrence, M. S., Voet, D., Jing, R., Cibulskis, K., Sivachenko, A., Stojanov, P., McKenna, A., Lander, E. S., Gabriel, S., Getz, G., Ding, L., Fulton, R. S., Koboldt, D. C., Wylie, T., Walker, J., Dooling, D. J., Fulton, L., Delehaunty, K. D., Fronick, C. C., Demeter, R., Mardis, E. R., Wilson, R. K., Chu, A. A. L., Chun, H.-J. E., Mungall, A. J., Pleasance, E., Gordon Robertson, A., Stoll, D., Balasundaram, M., Birol, I., Butterfield, Y. S. N., Chuah, E., Coope, R. J. N., Dhalla, N., Guin, R., Hirst, C., Hirst, M., Holt, R. A., Lee, D., Li, H. I., Mayo, M., Moore, R. A., Schein, J. E., Slobodan, J. R., Tam, A., Thiessen, N., Varhol, R., Zeng, T., Zhao, Y., Jones, S. J. M., Marra, M. A., Bass, A. J., Ramos, A. H., Saksena, G., Cherniack, A. D., Schumacher, S. E., Tabak, B., Carter, S. L., Pho, N. H., Nguyen, H., Onofrio, R. C., Crenshaw, A., Ardlie, K., Beroukhim, R., Winckler, W., Getz, G., Meyerson, M., Protopopov, A., Zhang, J., Hadjipanayis, A., Lee, E., Xi, R., Yang, L. L., Ren, X., Zhang, H., Sathiamoorthy, N., Shukla, S., Chen, P.-C., Haseley, P., Xiao, Y., Lee, S., Seidman, J., Chin, L., Park, P. J., Kucherlapati, R., Todd Auman, J., Hoadley, K. A., Du, Y., Wilkerson, M. D., Shi, Y., Liquori, C., Meng, S., Li, L., Turman, Y. J., Topal, M. D., Tan, D., Waring, S., Buda, E., Walsh, J., Jones, C. D., Mieczkowski, P. A., Singh, D., Wu, J., Gulabani, A., Dolina, P., Bodenheimer, T., Hoyle, A. P., Simons, J. V., Soloway, M., Mose, L. E., Jefferys, S. R., Balu, S., O'Connor, B. D., Prins, J. F., Chiang, D. Y., Neil Hayes, D., Perou, C. M., Hinoue, T., Weisenberger, D. J., Maglinte, D. T., Pan, F., Berman, B. P., Van Den Berg, D. J., Shen, H., Triche Jr, T., Baylin, S. B., Laird, P. W., Getz, G., Noble, M., Voet, D., Saksena, G., Gehlenborg, N., DiCara, D., Zhang, J., Zhang, H., Wu, C.-J., Yingchun Liu, S., Shukla, S., Lawrence, M. S., Zhou, L., Sivachenko, A., Lin, P., Stojanov, P., Jing, R., Park, R. W., Nazaire, M.-D., Robinson, J., Thorvaldsdottir, H., Mesirov, J., Park, P. J., Chin, L., Thorsson, V., Reynolds, S. M., Bernard, B., Kreisberg, R., Lin, J., Iype, L., Bressler, R., Erkkilä, T., Gundapuneni, M., Liu, Y., Norberg, A., Robinson, T., Yang, D., Zhang, W., Shmulevich, I., de Ronde, J. J., Schultz, N., Cerami, E., Ciriello, G., Goldberg, A. P., Gross, B., Jacobsen, A., Gao, J., Kaczkowski, B., Sinha, R., Arman Aksoy, B., Antipin, Y., Reva, B., Shen, R., Taylor, B. S., Chan, T. A., Ladanyi, M., Sander, C., Akbani, R., Zhang, N., Broom, B. M., Casasent, T., Unruh, A., Wakefield, C., Hamilton, S. R., Craig Cason, R., Baggerly, K. A., Weinstein, J. N., Haussler, D., Benz, C. C., Stuart, J. M., Benz, S. C., Zachary Sanborn, J., Vaske, C. J., Zhu, J., Szeto, C., Scott, G. K., Yau, C., Ng, S., Goldstein, T., Ellrott, K., Collisson, E., Cozen, A. E., Zerbino, D., Wilks, C., Craft, B., Spellman, P., Penny, R., Shelton, T., Hatfield, M., Morris, S., Yena, P., Shelton, C., Sherman, M., Paulauskis, J., Gastier-Foster, J. M., Bowen, J., Ramirez, N. C., Black, A., Pyatt, R., Wise, L., White, P., Bertagnolli, M., Brown, J., Chan, T. A., Chu, G. C., Czerwinski, C., Denstman, F., Dhir, R., Dörner, A., Fuchs, C. S., Guillem, J. G., Iacocca, M., Juhl, H., Kaufman, A., Kohl III, B., Van Le, X., Mariano, M. C., Medina, E. N., Meyers, M., Nash, G. M., Paty, P. B., Petrelli, N., Rabeno, B., Richards, W. G., Solit, D., Swanson, P., Temple, L., Tepper, J. E., Thorp, R., Vakiani, E., Weiser, M. R., Willis, J. E., Witkin, G., Zeng, Z., Zinner, M. J., Zornig, C., Jensen, M. A., Sfeir, R., Kahn, A. B., Chu, A. A. L., Kothiyal, P., Wang, Z., Snyder, E. E., Pontius, J., Pihl, T. D., Ayala, B., Backus, M., Walton, J., Whitmore, J., Baboud, J., Berton, D. L., Nicholls, M. C., Srinivasan, D., Raman, R., Girshik, S., Kigonya, P. A., Alonso, S., Sanbhadti, R. N., Barletta, S. P., Greene, J. M., Pot, D. A., Mills Shaw, K. R., Dillon, L. A. L., Buetow, K., Davidsen, T., Demchok, J. A., Eley, G., Ferguson, M., Fielding, P., Schaefer, C., Sheth, M., Yang, L. L., Guyer, M. S., Ozenberger, B. A., Palchik, J. D., Peterson, J., Sofia, H. J., and Thomson., E. (2012). Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407):330–7.

Myers, S., Bottolo, L., Freeman, C., McVean, G., and Donnelly, P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. *Science*, 310(5746):321–4.

Neph, S., Kuehn, M. S., Reynolds, A. P., Haugen, E., Thurman, R. E., Johnson, A. K., Rynes, E., Maurano, M. T., Vierstra, J., Thomas, S., Sandstrom, R., Humbert, R., and Stamatoyannopoulos, J. A. (2012). BEDOPS: high-performance genomic feature operations. *Bioinformatics*, 28(14):1919–20.

Newcomb, P. A., Baron, J., Cotterchio, M., Gallinger, S., Grove, J., Haile, R., Hall, D., Hopper, J. L., Jass, J., Le Marchand, L., Limburg, P., Lindor, N., Potter, J. D., Templeton, A. S., Thibodeau, S., and Seminara, D. (2007). Colon cancer family registry: An international resource for studies of the genetic epidemiology of colon cancer. *Cancer Epidemiology Biomarkers and Prevention*, 16(11):2331–2343.

Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Dolan, M. E., and Cox, N. J. (2010). Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS genetics*, 6(4):e1000888.

Nicolaides, N. C., Papadopoulos, N., Liu, B., Wei, Y. F., Carter, K. C., Ruben, S. M., Rosen, C. A., Haseltine, W. A., Fleischmann, R. D., and Fraser, C. M. (1994). Mutations of two PMS homologues in hereditary nonpolyposis colon cancer. *Nature*, 371(6492):75–80.

Niessen, R. C., Sijmons, R. H., Ou, J., Olthof, S. G. M., Osinga, J., Ligtenberg, M. J., Hogervorst, F. B. L., Weiss, M. M., Tops, C. M. J., Hes, F. J., de Bock, G. H., Buys, C. H. C. M., Kleibeuker, J. H., and Hofstra, R. M. W. (2006). MUTYH and the mismatch repair system: Partners in crime? *Human Genetics*, 119(1-2):206–211.

Nolan, K., Kattamuri, C., Rankin, S. A. A., Read, R. J. J., Zorn, A. M. M., and Thompson, T. B. B. (2016). Structure of Gremlin-2 in Complex with GDF5 Gives Insight into DAN-Family-Mediated BMP Antagonism. *Cell Reports*, 16(8):2077–86.

Nolan, K. and Thompson, T. B. (2014). The DAN family: modulators of TGF-β signaling and beyond. *Protein science*, 23(8):999–1012.

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., Indap, A., King, K. S., Bergmann, S., Nelson, M. R., Stephens, M., and Bustamante, C. D. (2008). Genes mirror geography within Europe. *Nature*, 456(7218):98–101.

Oetting, W. S. and King, R. A. (1993). Molecular basis of type I (tyrosinase-related) oculocutaneous albinism: Mutations and polymorphisms of the human tyrosinase gene. *Human Mutation*, 2(1):1–6.

Okabayashi, K., Ashrafian, H., Hasegawa, H., Yoo, J.-H., Patel, V. M., Harling, L., Rowland, S. P., Ali, M., Kitagawa, Y., Darzi, A., and Athanasiou, T. (2012). Body mass index category as a risk factor for colorectal adenomas: a systematic review and meta-analysis. *The American journal of gastroenterology*, 107(8):1175–85; quiz 1186.

Orlando, G., Law, P. J., Palin, K., Tuupanen, S., Gylfe, A., Hänninen, U. A., Cajuso, T., Tanskanen, T., Kondelin, J., Kaasinen, E., Sarin, A.-P., Kaprio, J., Eriksson, J. G., Rissanen, H., Knekt, P., Pukkala, E., Jousilahti, P., Salomaa, V., Ripatti, S., Palotie, A., Järvinen, H., Renkonen-Sinisalo, L., Lepistö, A., Böhm, J., Mecklin, J.-P., Al-Tassan, N. A., Palles, C., Martin, L., Barclay, E., Tenesa, A., Farrington, S., Timofeeva, M. N., Meyer, B. F., Wakil, S. M., Campbell, H., Smith, C. G., Idziaszczyk, S., Maughan, T. S., Kaplan, R., Kerr, R., Kerr, D., Buchanan, D. D., Ko Win, A., Hopper, J., Jenkins, M., Lindor, N. M., Newcomb, P. A., Gallinger, S., Conti, D., Schumacher, F., Casey, G., Taipale, J., Cheadle, J. P., Dunlop, M. G., Tomlinson, I. P., Aaltonen, L. A., and Houlston, R. S. (2016). Variation at 2q35 (PNKD and TMBIM1) influences colorectal cancer risk and identifies a pleiotropic effect with inflammatory bowel disease. *Human molecular genetics*, 25(11):2349–2359.

Østergaard, J. R., Sunde, L., and Okkels, H. (2005). Neurofibromatosis von Recklinghausen type I phenotype and early onset of cancers in siblings compound heterozygous for mutations inMSH6. *American Journal of Medical Genetics Part A*, 139A(2):96–105.

Ozaki, K., Ohnishi, Y., Iida, A., Sekine, A., Yamada, R., Tsunoda, T., Sato, H., Sato, H., Hori, M., Nakamura, Y., and Tanaka, T. (2002). Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. *Nature genetics*, 32(4):650–4.

Pacifico, L., Carducci, C., Poggiogalle, E., Caravona, F., Antonozzi, I., Chiesa, C., and Maggiore, G. (2010). Mutational analysis of ABCC2 gene in two siblings with neonatal-onset Dubin Johnson syndrome. *Clinical Genetics*, 78(6):598–600.

Palles, C., Cazier, J.-B., Howarth, K. M., Domingo, E., Jones, A. M., Broderick, P., Kemp, Z., Spain, S. L., Guarino, E., Guarino Almeida, E., Salguero, I., Sherborne, A., Chubb, D., Carvajal-Carmona, L. G., Ma, Y., Kaur, K., Dobbins, S., Barclay, E., Gorman, M., Martin, L., Kovac, M. B., Humphray, S., Lucassen, A., Holmes, C. C., Bentley, D., Donnelly, P., Taylor, J., Petridis, C., Roylance, R., Sawyer, E. J., Kerr, D. J., Clark, S., Grimes, J., Kearsey, S. E., Thomas, H. J. W., McVean, G., Houlston, R. S., and Tomlinson, I. (2013). Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. *Nature genetics*, 45(2):136–44.

Pande, M., Chen, J., Amos, C. I., Lynch, P. M., Broaddus, R., and Frazier, M. L. (2007). Influence of methylenetetrahydrofolate reductase gene polymorphisms C677T and A1298C on age-associated risk for colorectal cancer in a Caucasian Lynch syndrome population. *Cancer Epidemiology Biomarkers and Prevention*, 16(9):1753–1759.

Papaemmanuil, E., Carvajal-Carmona, L., Sellick, G. S., Kemp, Z., Webb, E., Spain, S., Sullivan, K., Barclay, E., Lubbe, S., Jaeger, E., Vijayakrishnan, J., Broderick, P., Gorman, M., Martin, L., Lucassen, A., Bishop, D. T., Evans, D. G., Maher, E. R., Steinke, V., Rahner, N., Schackert, H. K., Goecke, T. O., Holinski-Feder, E., Propping, P., Van Wezel, T., Wijnen, J., Cazier, J.-B., Thomas, H., Houlston, R. S., and Tomlinson, I. (2008). Deciphering the genetics of hereditary non-syndromic colorectal cancer. *European journal of human genetics : EJHG*, 16(12):1477–86.

Patnaik, P. B. (1949). The Non-Central χ2- and F-Distribution and their Applications. *Biometrika*, 36(1/2):202–232.

Payne, S. R., Church, T. R., Wandell, M., Rösch, T., Osborn, N., Snover, D., Day, R. W., Ransohoff, D. F., and Rex, D. K. (2014). Endoscopic detection of proximal serrated lesions and pathologic identification of sessile serrated adenomas/polyps vary on the basis of center. *Clinical Gastroenterology and Hepatology*, 12(7):1119–1126.

Pelli, A., Väyrynen, J. P., Klintrup, K., Mäkelä, J., Mäkinen, M. J., Tuomisto, A., and Karttunen, T. J. (2016). Gremlin1 expression associates with serrated pathway and favorable prognosis in colorectal cancer. *Histopathology*.

Peltomäki, P., Lothe, R. A., Aaltonen, L. A., Pylkkänen, L., Nyström-Lahti, M., Seruca, R., David, L., Holm, R., Ryberg, D., and Haugen, A. (1993). Microsatellite instability is associated with tumors that characterize the hereditary non-polyposis colorectal carcinoma syndrome. *Cancer research*, 53(24):5853–5.

Pendlebury, S., Duchesne, F., Reed, K. A., Smith, J. L., and Kerr, D. J. (2003). A trial of adjuvant therapy in colorectal cancer: the VICTOR trial. *Clinical colorectal cancer*, 3(1):58–60.

Perez-Villamil, B., Romera-Lopez, A., Hernandez-Prieto, S., Lopez-Campos, G., Calles, A., Lopez-Asenjo, J.-A., Sanz-Ortega, J., Fernandez-Perez, C., Sastre, J., Alfonso, R., Caldes, T., Martin-Sanchez, F., and Diaz-Rubio, E. (2012). Colon cancer molecular subtypes identified by expression profiling and associated to stroma, mucinous type and different clinical behavior.

Phipps, A. I., Limburg, P. J., Baron, J. A., Burnett-Hartman, A. N., Weisenberger, D. J., Laird, P. W., Sinicrope, F. A., Rosty, C., Buchanan, D. D., Potter, J. D., and Newcomb, P. A. (2014). Association Between Molecular Subtypes of Colorectal Cancer and Patient Survival. *Gastroenterology*.

Plaschke, J., Preußler, M., Ziegler, A., and Schackert, H. K. (2012). Aberrant protein expression and frequent allelic loss of MSH3 in colorectal cancer with low-level microsatellite instability. *Int J Colorectal Dis*, 27(7):911–919.

Plesec, T., Brown, K., Allen, C., Burke, C., Church, J., Kalady, M., LaGuardia, L., O'Malley, M., and Heald, B. (2016). Clinicopathologic Features of a Kindred with SCG5-GREM1-associated Hereditary Mixed Polyposis Syndrome. *Human Pathology*.

Polderman, T. J. C., Benyamin, B., de Leeuw, C. A., Sullivan, P. F., van Bochoven, A., Visscher, P. M., and Posthuma, D. (2015). Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature Genetics*, 47(7):702–709.

Power, C. and Elliott, J. (2006). Cohort profile: 1958 British birth cohort (National Child Development Study). *International journal of epidemiology*, 35(1):34–41.

Pregizer, S. and Mortlock, D. P. (2009). Control of BMP gene expression by long-range regulatory elements. *Cytokine & growth factor reviews*, 20(5-6):509–15.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909.

Price, A. L., Weale, M. E., Patterson, N., Myers, S. R., Need, A. C., Shianna, K. V., Ge, D., Rotter, J. I., Torres, E., Taylor, K. D. D., Goldstein, D. B., and Reich, D. (2008). Long-Range LD Can Confound Genome Scans in Admixed Populations. *American Journal of Human Genetics*, 83(1):132–135.

Pritchard, J. K. (2001). Are rare variants responsible for susceptibility to complex diseases? *American journal of human genetics*, 69(1):124–137.

Pruim, R. J., Welch, R. P., Sanna, S., Teslovich, T. M., Chines, P. S., Gliedt, T. P., Boehnke, M., Abecasis, G. R., and Willer, C. J. (2010). LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics*, 26(18):2336–7.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I., Daly, M. J., and Sham, P. C. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, 81(3):559–575.

Pursell, Z. F., Isoz, I., Lundström, E.-B., Johansson, E., and Kunkel, T. A. (2007). Yeast DNA polymerase epsilon participates in leading-strand DNA replication. *Science*, 317(5834):127–30.

Qi, Z., Li, Y., Zhao, B., Xu, C., Liu, Y., Li, H., Zhang, B., Wang, X., Yang, X., Xie, W., Li, B., Han, J.-D. J., and Chen, Y.-G. (2017). BMP restricts stemness of intestinal Lgr5+ stem cells by directly suppressing their signature genes. *Nature Communications*, 8:13824.

Quinlan, A. R. and Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842.

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Raczy, C., Petrovski, R., Saunders, C. T., Chorny, I., Kruglyak, S., Margulies, E. H., Chuang, H.-Y., Källberg, M., Kumar, S. A., Liao, A., Little, K. M., Strömberg, M. P., and Tanner, S. W. (2013). Isaac: ultra-fast whole-genome secondary analysis on Illumina sequencing platforms. *Bioinformatics*, 29(16):2041–3.

Rahib, L., Smith, B. D., Aizenberg, R., Rosenzweig, A. B., Fleshman, J. M., and Matrisian, L. M. (2014). Projecting cancer incidence and deaths to 2030: The unexpected burden of thyroid, liver, and pancreas cancers in the united states. *Cancer Research*, 74(11):2913–2921.

Rahmeh, A. A., Zhou, Y., Xie, B., Li, H., Lee, E. Y. C., and Lee, M. Y. W. T. (2012). Phosphorylation of the p68 subunit of Pol δ acts as a molecular switch to regulate its interaction with PCNA. *Biochemistry*, 51(1):416–24.

Ramachandran, S., Deshpande, O., Roseman, C. C., Rosenberg, N. A., Feldman, M. W., and Cavalli-Sforza, L. L. (2005). Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proceedings of the National Academy of Sciences of the United States of America*, 102(44):15942–15947.

Reeves, S. G., Meldrum, C., Groombridge, C., Spigelman, A. D., Suchy, J., Kurzawski, G., Lubinski, J., McElduff, P., and Scott, R. J. (2009). MTHFR 677 C>T and 1298 A>C polymorphisms and the age of onset of colorectal cancer in hereditary nonpolyposis colorectal cancer. *European journal of human genetics : EJHG*, 17(5):629–35.

Reeves, S. G., Rich, D., Meldrum, C. J., Colyvas, K., Kurzawski, G., Suchy, J., Lubinski, J., and Scott, R. J. (2008). IGF1 is a modifier of disease risk in hereditary non-polyposis colorectal cancer. *International Journal of Cancer*, 123(6):1339–1343.

Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J., Lavery, T., Kouyoumjian, R., Farhadian, S. F., Ward, R., and Lander, E. S. (2001). Linkage disequilibrium in the human genome. *Nature*, 411(6834):199–204.

Renfro, L. A., Grothey, A., Kerr, D., Haller, D. G., André, T., Van Cutsem, E., Saltz, L., Labianca, R., Loprinzi, C. L., Alberts, S. R., Schmoll, H., Twelves, C., Yothers, G., Sargent, D. J., Sargent, D. J., Green, E., Grothey, A., Alberts, S. R., Shi, Q., Renfro, L. A., Yothers, G., O'Connell, M. J., Wolmark, N., de Gramont, A., Gray, R., Kerr, D., Haller, D. G., Guthrie, K., Buyse, M., Labianca, R., Seitz, J. F., O'Callaghan, C. J., Francini, G., Catalano, P. J., Blanke, C. D., Andre, T., Goldberg, R. M., Benson, A., Twelves, C., Sirzen, F., Cisar, L., Cutsem, E. V., and Saltz, L. (2015). Survival following early-stage colon cancer: An ACCENT-based comparison of patients versus a matched international general population. *Annals of Oncology*, 26(5):950–958.

Rimm, E. B., Stampfer, M. J., Colditz, G. A., Chute, C. G., Litin, L. B., and Willett, W. C. (1990). Validity of self-reported waist and hip circumferences in men and women. *Epidemiology*, 1(6):466–473.

Rimmer, A., Phan, H., Mathieson, I., Iqbal, Z., Twigg, S. R. F., Wilkie, A. O. M., McVean, G., and Lunter, G. (2014). Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature genetics*, 46(8):912–8.

Risinger, J. I., Umar, A., Boyd, J., Berchuck, A., Kunkel, T. A., and Barrett, J. C. (1996). Mutation of MSH3 in endometrial cancer and evidence for its functional role in heteroduplex repair. *Nature genetics*, 14(1):102–5.

Roa, B. B., Garcia, C. A., Pentao, L., Killian, J. M., Trask, B. J., Suter, U., Snipes, G. J., Ortiz-Lopez, R., Shooter, E. M., Patel, P. I., and Lupski, J. R. (1993). Evidence for a recessive PMP22 point mutation in Charcot-Marie-Tooth disease type 1A. *Nature genetics*, 5(2):189–194.

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., and Mesirov, J. P. (2011). Integrative genomics viewer. *Nature Biotechnology*, 29(1):24–26.

Rodríguez-Soler, M., Pérez-Carbonell, L., Guarinos, C., Zapater, P., Castillejo, A., Barberá, V. M., Juárez, M., Bessa, X., Xicola, R. M., Clofent, J., Bujanda, L., Balaguer, F., Reñé, J. M., De-Castro, L., Marín-Gabriel, J. C., Lanas, A., Cubiella, J., Nicolás-Pérez, D., Brea-Fernández, A., Castellví-Bel, S., Alenda, C., Ruiz-Ponte, C., Carracedo, A., Castells, A., Andreu, M., Llor, X., Soto, J. L., Payá, A., and Jover, R. (2013). Risk of cancer in cases of suspected lynch syndrome without germline mutation. *Gastroenterology*, 144(5):926–932.e1.

Roepman, P., Schlicker, A., Tabernero, J., Majewski, I., Tian, S., Moreno, V., Snel, M. H., Chresta, C. M., Rosenberg, R., Nitsche, U., Macarulla, T., Capella, G., Salazar, R., Orphanides, G., Wessels, L. F. A., Bernards, R., and Simon, I. M. (2013). Colorectal cancer intrinsic subtypes predict chemotherapy benefit, deficient mismatch repair and epithelial-to-mesenchymal transition. *International Journal of Cancer*, 134(3):552–562.

Rohlin, A., Eiengård, F., Lundstam, U., Zagoras, T., Nilsson, S., Edsjö, A., Pedersen, J., Svensson, J., Skullman, S., Karlsson, B. G., Björk, J., and Nordling, M. (2015). GREM1 and POLE variants in hereditary colorectal cancer syndromes. *Genes, chromosomes & cancer*.

Roig, A. I., Eskiocak, U., Hight, S. K., Kim, S. B., Delgado, O., Souza, R. F., Spechler, S. J., Wright, W. E., and Shay, J. W. (2010). Immortalized epithelial cells derived from human colon biopsies express stem cell markers and differentiate in vitro. *Gastroenterology*, 138(3):1012–21.e1–5.

Rosmarin, D., Palles, C., Church, D., Domingo, E., Jones, A., Johnstone, E., Wang, H., Love, S., Julier, P., Scudder, C., Nicholson, G., Gonzalez-Neira, A., Martin, M., Sargent, D., Green, E., McLeod, H., Zanger, U. M., Schwab, M., Braun, M., Seymour, M., Thompson, L., Lacas, B., Boige, V., Ribelles, N., Afzal, S., Enghusen, H., Jensen, S. A., Etienne-Grimaldi, M. C., Milano, G., Wadelius, M., Glimelius, B., Garmo, H., Gusella, M., Lecomte, T., Laurent-Puig, P., Martinez-Balibrea, E., Sharma, R., Garcia-Foncillas, J., Kleibl, Z., Morel, A., Pignon, J. P., Midgley, R., Kerr, D., and

Tomlinson, I. (2014). Genetic markers of toxicity from capecitabine and other fluorouracil-based regimens: Investigation in the QUASAR2 study, systematic review, and meta-analysis. *Journal of Clinical Oncology*, 32(10):1031–1039.

Rosmarin, D., Palles, C., Pagnamenta, A., Kaur, K., Pita, G., Martin, M., Domingo, E., Jones, A., Howarth, K., Freeman-Mills, L., Johnstone, E., Wang, H., Love, S., Scudder, C., Julier, P., Fernández-Rozadilla, C., Ruiz-Ponte, C., Carracedo, A., Castellvi-Bel, S., Castells, A., Gonzalez-Neira, A., Taylor, J., Kerr, R., Kerr, D., and Tomlinson, I. (2015). A candidate gene study of capecitabine-related toxicity in colorectal cancer identifies new toxicity variants at DPYD and a putative role for ENOSF1 rather than TYMS. *Gut*, 64(1):111–120.

Rothwell, P. M., Wilson, M., Elwin, C. E., Norrving, B., Algra, A., Warlow, C. P., and Meade, T. W. (2010). Long-term effect of aspirin on colorectal cancer incidence and mortality: 20-year follow-up of five randomised trials. *The Lancet*, 376(9754):1741–1750.

Rustgi, A. K. (2007). The genetics of hereditary colon cancer. *Genes & development*, 21(20):2525–38.

Ryan, E., O'keane, C., and Crowe, J. (1998). Hemochromatosis in Ireland and HFE. *Blood cells, molecules & diseases*, 24(4):428–32.

Sadanandam, A., Lyssiotis, C. A., Homicsko, K., Collisson, E. A., Gibb, W. J., Wullschleger, S., Ostos, L. C. G., Lannon, W. A., Grotzinger, C., Del Rio, M., Lhermitte, B., Olshen, A. B., Wiedenmann, B., Cantley, L. C., Gray, J. W., and Hanahan, D. (2013). A colorectal cancer classification system that associates cellular phenotype and responses to therapy. *Nature medicine*, 19(5):619–25.

Samowitz, W. S., Albertsen, H., Herrick, J., Levin, T. R., Sweeney, C., Murtaugh, M. A., Wolff, R. K., and Slattery, M. L. (2005). Evaluation of a large, population-based sample supports a CpG island methylator phenotype in colon cancer. *Gastroenterology*, 129(3):837–845.

Sampson, J. N., Wheeler, W. A., Yeager, M., Panagiotou, O., Wang, Z., Berndt, S. I., Lan, Q., Abnet, C. C., Amundadottir, L. T., Figueroa, J. D., Landi, M. T., Mirabello, L., Savage, S. A., Taylor, P. R., Vivo, I. D., McGlynn, K. A., Purdue, M. P., Rajaraman, P., Adami, H.-O., Ahlbom, A., Albanes, D., Amary, M. F., An, S.-J., Andersson, U., Andriole, G., Andrulis, I. L., Angelucci, E., Ansell, S. M., Arici, C., Armstrong, B. K., Arslan, A. A., Austin, M. A., Baris, D., Barkauskas, D. A., Bassig, B. A., Becker, N., Benavente, Y., Benhamou, S., Berg, C., Van Den Berg, D., Bernstein, L., Bertrand, K. A., Birmann, B. M., Black, A., Boeing, H., Boffetta, P., Boutron-Ruault, M.-C., Bracci, P. M., Brinton, L., Brooks-Wilson, A. R., Bueno-de Mesquita, H. B., Burdett, L., Buring, J., Butler, M. A., Cai, Q., Cancel-Tassin, G., Canzian, F., Carrato, A., Carreon, T., Carta, A., Chan, J. K. C., Chang, E. T., Chang, G.-C., Chang, I.-S., Chang, J., Chang-Claude, J., Chen, C.-J., Chen, C.-Y., Chen, C., Chen, C.-H., Chen, C., Chen, H., Chen, K., Chen, K.-Y., Chen, K.-C., Chen, Y., Chen, Y.-H., Chen, Y.-S., Chen, Y.-M., Chien, L.-H., Chirlaque, M.-D., Choi, J. E., Choi, Y. Y., Chow, W.-H., Chung, C. C., Clavel, J., Clavel-Chapelon, F., Cocco, P., Colt, J. S., Comperat, E., Conde, L., Connors, J. M., Conti, D., Cortessis, V. K., Cotterchio, M., Cozen, W., Crouch, S., Crous-Bou, M., Cussenot, O., Davis, F. G., Ding, T., Diver, W. R., Dorronsoro, M., Dossus, L., Duell, E. J., Ennas, M. G., Erickson, R. L., Feychting, M., Flanagan, A. M., Foretova, L., Fraumeni, J. F., Freedman, N. D., Beane Freeman, L. E., Fuchs, C., Gago-Dominguez, M., Gallinger, S., Gao, Y.-T., Gapstur, S. M., Garcia-Closas, M., García-Closas, R., Gascoyne, R. D., Gastier-Foster, J., Gaudet, M. M., Gaziano, J. M., Giffen, C., Giles, G. G., Giovannucci, E., Glimelius, B., Goggins, M., Gokgoz, N., Goldstein, A. M., Gorlick, R., Gross, M., Grubb, R., Gu, J., Guan, P., Gunter, M., Guo, H., Habermann, T. M., Haiman, C. A., Halai, D., Hallmans, G., Hassan, M., Hattinger, C., He, Q., He, X., Helzlsouer, K., Henderson, B., Henriksson, R., Hjalgrim, H., Hoffman-Bolton, J., Hohensee, C., Holford, T. R., Holly, E. A., Hong, Y.-C., Hoover, R. N., Horn-Ross, P. L., Hosain, G. M. M., Hosgood, H. D., Hsiao, C.-F., Hu, N., Hu, W., Hu, Z., Huang, M.-S., Huerta, J.-M., Hung, J.-Y., Hutchinson, A., Inskip, P. D., Jackson, R. D., Jacobs, E. J., Jenab, M., Jeon, H.-S., Ji, B.-T., Jin, G., Jin, L., Johansen, C., Johnson, A., Jung, Y. J., Kaaks, R., Kamineni, A., Kane, E., Kang, C. H., Karagas, M. R., Kelly, R. S., Khaw, K.-T., Kim, C., Kim, H. N., Kim, J. H., Kim, J. S., Kim, Y. H., Kim, Y. T., Kim, Y.-C., Kitahara, C. M., Klein, A. P., Klein, R. J., Kogevinas, M., Kohno, T., Kolonel, L. N., Kooperberg, C., Kricker, A., Krogh, V., Kunitoh, H., Kurtz, R. C., Kweon, S.-S., LaCroix, A., Lawrence, C., Lecanda, F., Lee, V. H. F., Li, D., Li, H., Li, J., Li, Y.-J., Li, Y., Liao, L. M., Liebow, M., Lightfoot, T., Lim, W.-Y., Lin, C.-C., Lin, D., Lindstrom, S., Linet, M. S., Link, B. K., Liu, C., Liu, J., Liu, L., Ljungberg, B., Lloreta, J., Lollo, S. D., Lu, D., Lund, E., Malats, N., Mannisto, S., Marchand, L. L., Marina, N., Masala, G., Mastrangelo, G., Matsuo, K., Maynadie, M., McKay, J., McKean-Cowdin,

R., Melbye, M., Melin, B. S., Michaud, D. S., Mitsudomi, T., Monnereau, A., Montalvan, R., Moore, L. E., Mortensen, L. M., Nieters, A., North, K. E., Novak, A. J., Oberg, A. L., Offit, K., Oh, I.-J., Olson, S. H., Palli, D., Pao, W., Park, I. K., Park, J. Y., Park, K. H., Patiño-Garcia, A., Pavanello, S., Peeters, P. H. M., Perng, R.-P., Peters, U., Petersen, G. M., Picci, P., Pike, M. C., Porru, S., Prescott, J., Prokunina-Olsson, L., Qian, B., Qiao, Y.-L., Rais, M., Riboli, E., Riby, J., Risch, H. A., Rizzato, C., Rodabough, R., Roman, E., Roupret, M., Ruder, A. M., de Sanjose, S., Scelo, G., Schned, A., Schumacher, F., Schwartz, K., Schwenn, M., Scotlandi, K., Seow, A., Serra, C., Serra, M., Sesso, H. D., Setiawan, V. W., Severi, G., Severson, R. K., Shanafelt, T. D., Shen, H., Shen, W., Shin, M.-H., Shiraishi, K., Shu, X.-O., Siddiq, A., Sierrasesúmaga, L., Sihoe, A. D. L., Skibola, C. F., Smith, A., Smith, M. T., Southey, M. C., Spinelli, J. J., Staines, A., Stampfer, M., Stern, M. C., Stevens, V. L., Stolzenberg-Solomon, R. S., Su, J., Su, W.-C., Sund, M., Sung, J. S., Sung, S. W., Tan, W., Tang, W., Tardón, A., Thomas, D., Thompson, C. A., Tinker, L. F., Tirabosco, R., Tjønneland, A., Travis, R. C., Trichopoulos, D., Tsai, F.-Y., Tsai, Y.-H., Tucker, M., Turner, J., Vajdic, C. M., Vermeulen, R. C. H., Villano, D. J., Vineis, P., Virtamo, J., Visvanathan, K., Wactawski-Wende, J., Wang, C., Wang, C.-L., Wang, J.-C., Wang, J., Wei, F., Weiderpass, E., Weiner, G. J., Weinstein, S., Wentzensen, N., White, E., Witzig, T. E., Wolpin, B. M., Wong, M. P., Wu, C., Wu, G., Wu, J., Wu, T., Wu, W., Wu, X., Wu, Y.-L., Wunder, J. S., Xiang, Y.-B., Xu, J., Xu, P., (2015). Analysis of Heritability and Shared Heritability Based on Genome-Wide Association Studies for Thirteen Cancer Types. *Journal of the National Cancer Institute*, 107(12):djv279.

Schadt, E. E., Molony, C., Chudin, E., Hao, K., Yang, X., Lum, P. Y., Kasarskis, A., Zhang, B., Wang, S., Suver, C., Zhu, J., Millstein, J., Sieberts, S., Lamb, J., GuhaThakurta, D., Derry, J., Storey, J. D., Avila-Campillo, I., Kruger, M. J., Johnson, J. M., Rohl, C. A., Van Nas, A., Mehrabian, M., Drake, T. A., Lusis, A. J., Smith, R. C., Guengerich, F. P., Strom, S. C., Schuetz, E., Rushmore, T. H., and Ulrich, R. (2008). Mapping the genetic architecture of gene expression in human liver. *PLoS Biology*, 6(5):1020–1032.

Schlicker, A., Beran, G., Chresta, C. M., McWalter, G., Pritchard, A., Weston, S., Runswick, S., Davenport, S., Heathcote, K., Castro, D. A., Orphanides, G., French, T., Wessels, L. F., Jemal, A., Bray, F., Center, M., Ferlay, J., Ward, E., Forman, D., Tomeo, C., Colditz, G., Willett, W., Giovannucci, E., Platz, E., Rockhill, B., Dart, H., Hunter, D., Fearon, E., Saif, M., Chu, E., Issa, J.-P., Sanchez, J., Krumroy, L., Plummer, S., Aung, P., Merkulova, A., Skacel, M., DeJulius, K., Manilich, E., Church, J., Casey, G., Kalady, M., Markowitz, S., Bertagnolli, M., Iacopetta, B., Grieu, F., Amanuel, B., van Engeland, M., Derks, S., Smits, K., Meijer, G., Herman, J., Chibaudel, B., Tournigand, C., André, T., Larsen, A., de Gramont, A., Cutsem, E., Köhne, C.-H., Láng, I., Folprecht, G., Nowacki, M., Cascinu, S., Shchepotin, I., Maurel, J., Cunningham, D., Tejpar, S., Schlichting, M., Zubel, A., Celik, I., Rougier, P., Ciardiello, F., Molinari, F., Felicioni, L., Buscarino, M., Dosso, S., Buttitta, F., Malatesta, S., Movilia, A., Luoni, M., Boldorini, R., Alabiso, O., Girlando, S., Soini, B., Spitale, A., Nicolantonio, F., Saletti, P., Crippa, S., Mazzucchelli, L., Marchetti, A., Bardelli, A., Frattini, M., Sullivan, K., Kozuch, P., Prahallad, A., Sun, C., Huang, S., Nicolantonio, F. D., Salazar, R., Zecchin, D., Beijersbergen, R., Bardelli, A., Bernards, R., Dry, J., Pavey, S., Pratilas, C., Harbron, C., Runswick, S., Hodgson, D., Chresta, C., McCormack, R., Byrne, N., Cockerill, M., Graham, A., Beran, G., Cassidy, A., Haggerty, C., Brown, H., Ellison, G., Dering, J., Taylor, B., Stark, M., Bonazzi, V., Ravishankar, S., Packer, L., Xing, F., Solit, D., Finn, R., Rosen, N., Hayward, N., French, T., Smith, P., Garnett, M., Edelman, E., Heidorn, S., Greenman, C., Dastur, A., Lau, K., Greninger, P., Thompson, I., Luo, X., Soares, J., Liu, Q., Iorio, F., Surdez, D., Chen, L., Milano, R., Bignell, G., Tam, A., Davies, H., Stevenson, J., Barthorpe, S., Lutz, S., Kogera, F., Lawrence, K., McLaren-Douglas, A., Mitropoulos, X., Mironenko, T., Thi, H., Richardson, L., Zhou, W., Jewitt, F., Zhang, T., O'Brien, P., Boisvert, J., Price, S., Hur, W., Yang, W., Deng, X., Butler, A., Choi, H., Chang, J., Baselga, J., Stamenkovic, I., Engelman, J., Sharma, S., Delattre, O., Saez-Rodriguez, J., Gray, N., Settleman, J., Futreal, P., Haber, D., Stratton, M., Ramaswamy, S., McDermott, U., Benes, C., Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A., Kim, S., Wilson, C., Lehár, J., Kryukov, G., Sonkin, D., Reddy, A., Liu, M., Murray, L., Berger, M., Monahan, J., Morais, P., Meltzer, J., Korejwa, A., Jané-Valbuena, J., Mapa, F., Thibault, J., Bric-Furlong, E., Raman, P., Shipway, A., Engels, I., Cheng, J., Yu, G., Yu, J., Aspesi, P., Silva, M. D., Jagtap, K., Jones, M., Wang, L., Hatton, C., Palescandolo, E., Gupta, S., Mahan, S., Sougnez, C., Onofrio, R., Liefeld, T., MacConaill, L., Winckler, W., Reich, M., Li, N., Mesirov, J., Gabriel, S., Getz, G., Ardlie, K., Chan, V., Myer, V., Weber, B., Porter, J., Warmuth, M., Finan, P., Harris, J., Meyerson, M., Golub, T., Morrissey, M., Sellers, W., Schlegel, R., Garraway, L., Arango, D., Laiho, P., Kokko, A., Alhopuro, P., Sammalkorpi, H., Salovaara, R., Nicorici, D., Hautaniemi, S., Alazzouzi, H., Mecklin, J.-P., Järvinen, H., Hemminki, A., Astola, J., Schwartz, S., Aaltonen, L., Barrier, A., Boelle, P.-Y., Roser, F., Gregg,

J., Tse, C., Brault, D., Lacaine, F., Houry, S., Huguier, M., Franc, B., Flahault, A., Lemoine, A., Dudoit, S., Eschrich, S., Yang, I., Bloom, G., Kwong, K., Boulware, D., Cantor, A., Coppola, D., Kruhøffer, M., Aaltonen, L., Orntoft, T., Quackenbush, J., Yeatman, T., Garman, K., Acharya, C., Edelman, E., Grade, M., Gaedcke, J., Sud, S., Barry, W., Diehl, A., Provenzale, D., Ginsburg, G., Ghadimi, B., Ried, T., Nevins, J., Mukherjee, S., Hsu, D., Potti, A., Lin, Y.-H., Friederichs, J., Black, M., Mages, J., Rosenberg, R., Guilford, P., Phillips, V., Thompson-Fawcett, M., Kasabov, N., Toro, T., Merrie, A., van Rij, A., Yoon, H.-S., McCall, J., Siewert, J., Holzmann, B., Reeve, A., O'Connell, M., Lavery, I., Yothers, G., Paik, S., Clark-Langone, K., Lopatin, M., Watson, D., Baehner, F., Shak, S., Baker, J., Cowens, J., Wolmark, N., Salazar, R., Roepman, P., Capella, G., Moreno, V., Simon, I., Dreezen, C., Lopez-Doriga, A., Santos, C., Marijnen, C., Westerga, J., Bruin, S., Kerr, D., Kuppen, P., van de Velde, C., Morreau, H., Velthuysen, L., Glas, A., Veer, L., Tollenaar, R., Wang, Y., Jatkoe, T., Zhang, Y., Mutch, M., Talantov, D., Jiang, J., McLeod, H., Atkins, D., Jorissen, R., Gibbs, P., Christie, M., Prakash, S., Lipton, L., Desai, J., Kerr, D., Aaltonen, L., Arango, D., Kruhøffer, M., Orntoft, T., Andersen, C., Gruidl, M., Kamath, V., Eschrich, S., Yeatman, T., Sieber, O., Birkenkamp-Demtroder, K., Christensen, L., Olesen, S., Frederiksen, C., Laiho, P., Aaltonen, L., Laurberg, S., Sørensen, F., Hag (2012). Subtypes of primary colorectal tumors correlate with response to targeted treatment in colorectal cell lines. *BMC Medical Genomics*, 5(1):66.

Schumacher, F. R., Schmit, S. L., Jiao, S., Edlund, C. K., Wang, H., Zhang, B., Hsu, L., Huang, S.-C., Fischer, C. P., Harju, J. F., Idos, G. E., Lejbkowicz, F., Manion, F. J., McDonnell, K., McNeil, C. E., Melas, M., Rennert, H. S., Shi, W., Thomas, D. C., Van Den Berg, D. J., Hutter, C. M., Aragaki, A. K., Butterbach, K., Caan, B. J., Carlson, C. S., Chanock, S. J., Curtis, K. R., Fuchs, C. S., Gala, M., Giocannucci, E. L., Gogarten, S. M., Hayes, R. B., Henderson, B., Hunter, D. J., Jackson, R. D., Kolonel, L. N., Kooperberg, C., Kury, S., LaCroix, A., Laurie, C. C. A., Laurie, C. C. A., Lemire, M., Levine, D., Ma, J., Makar, K. W., Qu, C., Taverna, D., Ulrich, C. M., Wu, K., Kono, S., West, D. W., Berndt, S. I., Bezieau, S., Brenner, H., Campbell, P. T., Chan, A. T., Chang-Claude, J., Coetzee, G. A., Conti, D. V., Duggan, D., Figueiredo, J. C., Fortini, B. K., Gallinger, S. J., Gauderman, W. J., Giles, G., Green, R., Haile, R., Harrison, T. A., Hoffmeister, M., Hopper, J. L., Hudson, T. J., Jacobs, E., Iwasaki, M., Jee, S. H., Jenkins, M., Jia, W.-H., Joshi, A., Li, L., Lindor, N. M., Matsuo, K., Moreno, V., Mukherjee, B., Newcomb, P. A., Potter, J. D., Raskin, L., Rennert, G., Rosse, S., Severi, G., Schoen, R. E., Seminara, D., Shu, X.-O., Slattery, M. L., Tsugane, S., White, E., Xiang, Y.-B., Zanke, B. W., Zheng, W., Le Marchand, L., Casey, G., Gruber, S. B., and Peters, U. (2015). Genome-wide association study of colorectal cancer identifies six new susceptibility loci. *Nature communications*, 6:7138.

Scott, I. C., Blitz, I. L., Pappano, W. N., Maas, S. A., Cho, K. W., and Greenspan, D. S. (2001). Homologues of Twisted gastrulation are extracellular cofactors in antagonism of BMP signalling. *Nature*, 410(6827):475–8.

Seguí, N., Mina, L. B., Lázaro, C., Sanz-Pamplona, R., Pons, T., Navarro, M., Bellido, F., López-Doriga, A., Valdés-Mas, R., Pineda, M., Guinó, E., Vidal, A., Soto, J. L., Caldés, T., Durrán, M., Urioste, M., Rueda, D., Brunet, J., Balbín, M., Blay, P., Iglesias, S., Garré, P., Lastra, E., Sánchez-Heras, A. B., Valencia, A., Moreno, V., Pujana, M. n., Villanueva, A., Blanco, I., Capellá, G., Surrallés, J., Puente, X. S., and Valle, L. (2015). Germline Mutations in FAN1 Cause Hereditary Colorectal Cancer by Impairing DNA Repair. *Gastroenterology*, 149(3):563–566.

Shi, Z., Johnstone, D., Talseth-Palmer, B. A., Evans, T. J., Spigelman, A. D., Groombridge, C., Milward, E. A., Olynyk, J. K., Suchy, J., Kurzawski, G., Lubinski, J., and Scott, R. J. (2009). Haemochromatosis HFE gene polymorphisms as potential modifiers of hereditary nonpolyposis colorectal cancer risk and onset age. *International Journal of Cancer*, 125(1):78–83.

Shifman, S., Kuypers, J., Kokoris, M., Yakir, B., and Darvasi, A. (2003). Linkage disequilibrium patterns of the human genome across populations. *Human Molecular Genetics*, 12(7):771–776.

Shim, H.-J., Lee, R., Shin, M.-H., Kim, H.-N., and Kweon, S.-S. (2016). Association between the TCF7L2 polymorphism and colorectal cancer does not differ by diabetes and obesity statuses. *Cancer Epidemiology*, 45:108–111.

Shin, J.-H., Blay, S., McNeney, B., and Graham, J. (2006). Ldheatmap: An r function for graphical display of pairwise linkage disequilibria between single nucleotide polymorphisms. *J Stat Soft*, 16:Code Snippet 3.

Short, E., Thomas, L. E., Hurley, J., Jose, S., and Sampson, J. R. (2015). Inherited predisposition to colorectal cancer: towards a more complete picture. *Journal of medical genetics*, 52(12):jmedgenet–2015–103298.

Sieber, O. M., Lipton, L., Crabtree, M., Heinimann, K., Fidalgo, P., Phillips, R. K. S., Bisgaard, M.-L., Orntoft, T. F., Aaltonen, L. A., Hodgson, S. V., Thomas, H. J. W., and Tomlinson, I. P. M. (2003). Multiple colorectal adenomas, classic adenomatous polyposis, and germ-line mutations in MYH. *The New England journal of medicine*, 348(9):791–9.

Siva, N. (2015). UK gears up to decode 100 000 genomes from NHS patients. *The Lancet*, 385(9963):103–104.

Slattery, M. L., Friedman, G. D., Potter, J. D., Edwards, S., Caan, B. J., and Samowitz, W. (1996). A description of age, sex, and site distributions of colon carcinoma in three geographic areas. *Cancer*, 78(8):1666–1670.

Smith, C. G., Naven, M., Harris, R., Colley, J., West, H., Li, N., Liu, Y., Adams, R., Maughan, T. S., Nichols, L., Kaplan, R., Wagner, M. J., Mcleod, H. L., and Cheadle, J. P. (2013). Exome Re-Sequencing Identifies Potential Tumor Suppressor Genes that Predispose to Colorectal Cancer. *Human Mutation*, 34(7):1026–1034.

Smith, R., Healy, E., Siddiqui, S., Flanagan, N., Steijlen, P. M., Rosdahl, I., Jacques, J. P., Rogers, S., Turner, R., Jackson, I. J., Birch-Machin, M. A., and Rees, J. L. (1998). Melanocortin 1 receptor variants in an Irish population. *Journal of Investigative Dermatology*, 111(1):119–122.

Snover, D., Ahnen, D., and Burt, R. (2010). Serrated polyps of the colon and rectum and serrated polyposis. In Bosman, F., Carneiro, F., Hruban, R., and Theise, N., editors, *WHO Classification of Tumours of the Digestive System*, pages 160–165. International Agency for Research on Cancer, Lyon, France, 4th edition.

Sobin, L. H. and Fleming, I. D. (1997). TNM classification of malignant tumors, Fifth edition (1997). *Cancer*, 80(9):1803–1804.

Solomon, E., Voss, R., Hall, V., Bodmer, W. F., Jass, J. R., Jeffreys, A. J., Lucibello, F. C., Patel, I., and Rider, S. H. (1987). Chromosome 5 allele loss in human colorectal carcinomas. *Nature*, 328(6131):616–9.

Spain, S. L., Carvajal-Carmona, L. G., Howarth, K. M., Jones, A. M., Su, Z., Cazier, J.-B., Williams, J., Aaltonen, L. A., Pharoah, P., Kerr, D. J., Cheadle, J., Li, L., Casey, G., Vodicka, P., Sieber, O., Lipton, L., Gibbs, P., Martin, N. G., Montgomery, G. W., Young, J., Baird, P. N., Morreau, H., van Wezel, T., Ruiz-Ponte, C., Fernandez-Rozadilla, C., Carracedo, A., Castells, A., Castellvi-Bel, S., Dunlop, M., Houlston, R. S., and Tomlinson, I. P. M. (2012). Refinement of the associations between risk of colorectal cancer and polymorphisms on chromosomes 1q41 and 12q13.13. *Human molecular genetics*, 21(4):934–46.

Srivatsan, A., Bowen, N., and Kolodner, R. D. (2014). Mispair-specific recruitment of the Mlh1-Pms1 complex identifies repair substrates of the Saccharomyces cerevisiae Msh2-Msh3 complex. *Journal of Biological Chemistry*, 289(13):9352–9364.

Stern, C. (1931). Zytologisch-genetische Untersuchungen als Beweise fur die Morgansche Theorie des Faktorenaustauschs. *Biol. Zentrablb.*, 51:547–587.

Sterne, J. A. and Egger, M. (2001). Funnel plots for detecting bias in meta-analysis. *Journal of Clinical Epidemiology*, 54(10):1046–1055.

Stewart, B. W. and Wild, C. P. (2014). World Cancer Report 2014. Technical report, World Health Organisation.

Strittmatter, W. J. (2012). Old Drug, New Hope for Alzheimer's Disease. *Science*, 335(6075):1447–1448.

Stunnenberg, H. G. and Hubner, N. C. (2014). Genomics meets proteomics: Identifying the culprits in disease. *Human Genetics*, 133(6):689–700.

Su, Z., Cardin, N., Donnelly, P., and Marchini, J. (2009). A Bayesian Method for Detecting and Characterizing Allelic Heterogeneity and Boosting Signals in Genome-Wide Association Studies. *Statistical Science*, 24(4):430–450.

Sulem, P., Gudbjartsson, D. F., Stacey, S. N., Helgason, A., Rafnar, T., Magnusson, K. P., Manolescu, A., Karason, A., Palsson, A., Thorleifsson, G., Jakobsdottir, M., Steinberg, S., Pálsson, S., Jonasson, F., Sigurgeirsson, B., Thorisdottir, K., Ragnarsson, R., Benediktsdottir, K. R., Aben, K. K., Kiemeney, L. A., Olafsson, J. H., Gulcher, J., Kong, A., Thorsteinsdottir, U., and Stefansson, K. (2007). Genetic determinants of hair, eye and skin pigmentation in Europeans. *Nature genetics*, 39(12):1443–52.

Sumi, S., Imaeda, M., Kidouchi, K., Ohba, S., Hamajima, N., Kodama, K., Togari, H., and Wada, Y. (1998). Population and family studies of dihydropyrimidinuria: Prevalence, inheretance mode and risk of fluorouracil toxicity. *American Journal of Medical Genetics*, 78(4):336–340.

Sutton, W. S. (1903). The Chromosomes in Heredity. *Biological Bulletin*, 4(5):231.

Tamborero, D., Gonzalez-Perez, A., and Lopez-Bigas, N. (2013). OncodriveCLUST: Exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics*, 29(18):2238–2244.

Tan, A., Abecasis, G. R., and Kang, H. M. (2015). Unified representation of genetic variants. *Bioinformatics*, 31(13):2202–2204.

Taupin, D., Lam, W., Rangiah, D., McCallum, L., Whittle, B., Zhang, Y., Andrews, D., Field, M., Goodnow, C. C., and Cook, M. C. (2015). A deleterious RNF43 germline mutation in a severely affected serrated polyposis kindred. *Human Genome Variation*, 2(February):15013.

Taylor, J. C., Martin, H. C., Lise, S., Broxholme, J., Cazier, J.-B., Rimmer, A., Kanapin, A., Lunter, G., Fiddy, S., Allan, C., Aricescu, A. R., Attar, M., Babbs, C., Becq, J., Beeson, D., Bento, C., Bignell, P., Blair, E., Buckle, V. J., Bull, K., Cais, O., Cario, H., Chapel, H., Copley, R. R., Cornall, R., Craft, J., Dahan, K., Davenport, E. E., Dendrou, C., Devuyst, O., Fenwick, A. L., Flint, J., Fugger, L., Gilbert, R. D., Goriely, A., Green, A., Greger, I. H., Grocock, R., Gruszczyk, A. V., Hastings, R., Hatton, E., Higgs, D., Hill, A., Holmes, C., Howard, M., Hughes, L., Humburg, P., Johnson, D., Karpe, F., Kingsbury, Z., Kini, U., Knight, J. C., Krohn, J., Lamble, S., Langman, C., Lonie, L., Luck, J., McCarthy, D., McGowan, S. J., McMullin, M. F., Miller, K. A., Murray, L., Németh, A. H., Nesbit, M. A., Nutt, D., Ormondroyd, E., Oturai, A. B., Pagnamenta, A., Patel, S. Y., Percy, M., Petousi, N., Piazza, P., Piret, S. E., Polanco-Echeverry, G., Popitsch, N., Powrie, F., Pugh, C., Quek, L., Robbins, P. A., Robson, K., Russo, A., Sahgal, N., van Schouwenburg, P. A., Schuh, A., Silverman, E., Simmons, A., Sørensen, P. S., Sweeney, E., Taylor, J., Thakker, R. V., Tomlinson, I., Trebes, A., Twigg, S. R. F., Uhlig, H. H., Vyas, P., Vyse, T., Wall, S. A., Watkins, H., Whyte, M. P., Witty, L., Wright, B., Yau, C., Buck, D., Humphray, S., Ratcliffe, P. J., Bell, J. I., Wilkie, A. O. M., Bentley, D., Donnelly, P., and McVean, G. (2015). Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. *Nature Genetics*, 47(7):717–726.

Tenesa, A., Farrington, S. M., Prendergast, J. G. D., Porteous, M. E., Walker, M., Haq, N., Barnetson, R. A., Theodoratou, E., Cetnarskyj, R., Cartwright, N., Semple, C., Clark, A. J., Reid, F. J. L., Smith, L. A., Kavoussanakis, K., Koessler, T., Pharoah, P. D. P., Buch, S., Schafmayer, C., Tepel, J., Schreiber, S., Völzke, H., Schmidt, C. O., Hampe, J., Chang-Claude, J., Hoffmeister, M., Brenner, H., Wilkening, S., Canzian, F., Capella, G., Moreno, V., Deary, I. J., Starr, J. M., Tomlinson, I. P. M., Kemp, Z., Howarth, K., Carvajal-Carmona, L., Webb, E., Broderick, P., Vijayakrishnan, J., Houlston, R. S., Rennert, G., Ballinger, D., Rozek, L., Gruber, S. B., Matsuda, K., Kidokoro, T., Nakamura, Y., Zanke, B. W., Greenwood, C. M. T., Rangrej, J., Kustra, R., Montpetit, A., Hudson, T. J., Gallinger, S., Campbell, H., and Dunlop, M. G. (2008). Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. *Nature genetics*, 40(5):631–7.

The ENCODE Project Consortium, Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., Epstein, C. B., Frietze, S., Harrow, J., Kaul, R., Khatun, J., Lajoie, B. R., Landt, S. G., Lee, B.-K., Pauli, F., Rosenbloom, K. R., Sabo, P., Safi, A., Sanyal, A., Shoresh, N., Simon, J. M., Song, L., Trinklein, N. D., Altshuler, R. C., Birney, E., Brown, J. B., Cheng, C., Djebali, S., Dong, X., Dunham, I., Ernst, J., Furey, T. S., Gerstein, M., Giardine, B., Greven, M., Hardison, R. C.,

Harris, R. S., Herrero, J., Hoffman, M. M., Iyer, S., Kellis, M., Khatun, J., Kheradpour, P., Kundaje, A., Lassmann, T., Li, Q., Lin, X., Marinov, G. K., Merkel, A., Mortazavi, A., Parker, S. C. J., Reddy, T. E., Rozowsky, J., Schlesinger, F., Thurman, R. E., Wang, J., Ward, L. D., Whitfield, T. W., Wilder, S. P., Wu, W., Xi, H. S., Yip, K. Y., Zhuang, J., Bernstein, B. E., Birney, E., Dunham, I., Green, E. D., Gunter, C., Snyder, M., Pazin, M. J., Lowdon, R. F., Dillon, L. A. L., Adams, L. B., Kelly, C. J., Zhang, J., Wexler, J. R., Green, E. D., Good, P. J., Feingold, E. A., Bernstein, B. E., Birney, E., Crawford, G. E., Dekker, J., Elnitski, L., Farnham, P. J., Gerstein, M., Giddings, M. C., Gingeras, T. R., Green, E. D., Guigó, R., Hardison, R. C., Hubbard, T. J., Kellis, M., Kent, W. J., Lieb, J. D., Margulies, E. H., Myers, R. M., Snyder, M., Stamatoyannopoulos, J. A., Tenenbaum, S. A., Weng, Z., White, K. P., Wold, B., Khatun, J., Yu, Y., Wrobel, J., Risk, B. A., Gunawardena, H. P., Kuiper, H. C., Maier, C. W., Xie, L., Chen, X., Giddings, M. C., Bernstein, B. E., Epstein, C. B., Shoresh, N., Ernst, J., Kheradpour, P., Mikkelsen, T. S., Gillespie, S., Goren, A., Ram, O., Zhang, X., Wang, L., Issner, R., Coyne, M. J., Durham, T., Ku, M., Truong, T., Ward, L. D., Altshuler, R. C., Eaton, M. L., Kellis, M., Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G. K., Khatun, J., Williams, B. A., Zaleski, C., Rozowsky, J., Röder, M., Kokocinski, F., Abdelhamid, R. F., Alioto, T., Antoshechkin, I., Baer, M. T., Batut, P., Bell, I., Bell, K., Chakrabortty, S., Chen, X., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M. J., Gao, H., Gonzalez, D., Gordon, A., Gunawardena, H. P., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Li, G., Luo, O. J., Park, E., Preall, J. B., Presaud, K., Ribeca, P., Risk, B. A., Robyr, D., Ruan, X., Sammeth, M., Sandhu, K. S., Schaeffer, L., See, L.-H., Shahab, A., Skancke, J., Suzuki, A. M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, H., Wrobel, J., Yu, Y., Hayashizaki, Y., Harrow, J., Gerstein, M., Hubbard, T. J., Reymond, A., Antonarakis, S. E., Hannon, G. J., Giddings, M. C., Ruan, Y., Wold, B., Carninci, P., Guigó, R., Gingeras, T. R., Rosenbloom, K. R., Sloan, C. A., Learned, K., Malladi, V. S., Wong, M. C., Barber, G. P., Cline, M. S., Dreszer, T. R., Heitner, S. G., Karolchik, D., Kent, W. J., Kirkup, V. M., Meyer, L. R., Long, J. C., Maddren, M., Raney, B. J., Furey, T. S., Song, L., Grasfeder, L. L., Giresi, P. G., Lee, B.-K., Battenhouse, A., Sheffield, N. C., Simon, J. M., Showers, K. A., Safi, A., London, D., Bhinge, A. A., Shestak, C., Schaner, M. R., Ki Kim, S., Zhang, Z. Z., Mieczkowski, P. A., Mieczkowska, J. O., Liu, Z., McDaniell, R. M., Ni, Y., Rashid, N. U., Kim, M. J., Adar, S., Zhang, Z., Wang, T., Winter, D., Keefe, D., Birney, E., Iyer, V. R., Lieb, J. D., Crawford, G. E., Li, G., Sandhu, K. S., Zheng, M., Wang, P., Luo, O. J., Shahab, A., Fullwood, M. J., Ruan, X., Ruan, Y., Myers, R. M., Pauli, F., Williams, B. A., Gertz, J., Marinov, G. K., Reddy, T. E., Vielmetter, J., Partridge, E., Trout, D., Varley, K. E., Gasper, C., Bansal, A., Pepke, S., Jain, P., Amrhein, H., Bowling, K. M., Anaya, M., Cross, M. K., King, B., Muratet, M. A., Antoshechkin, I., Newberry, K. M., McCue, K., Nesmith, A. S., Fisher-Aylor, K. I., Pusey, B., DeSalvo, G., Parker, S. L., Balasubramanian, S., Davis, N. S., Meadows, S. K., Eggleston, T., Gunter, C., Newberry, J. S., Levy, S. E., Absher, D. M., Mortazavi, A., Wong, W. H., Wold, B., Blow, M. J., Visel, A., Pennachio, L. A., Elnitski, L., Margulies, E. H., Parker, S. C. J., Petrykowska, H. M., Abyzov, A., Aken, B., Barrell, D., Barson, G., Berry, A., Bignell, A., Boychenko, V., Bussotti, G., Chrast, J., Davidson, C., Derrien, T., Despacio-Reyes, G., Diekhans, M., Ezkurdia, I., Frankish, A., Gilbert, J., Gonzalez, J. M., Griffiths, E., Harte, R., Hendrix, D. A., Howald, C., Hunt, T., Jungreis, I., Kay, M., Khurana, E., Kokocinski, F., Leng, J., Lin, M. F., Loveland, J., Lu, Z., Manthravadi, D., Mariotti, M., Mudge, J., Mukherjee, G., Notredame, C., Pei, B., Rodriguez, J. M., Saunders, G., Sboner, A., Searle, S., Sisu, C., Snow, C., Steward, C., Tanzer, A., Tapanari, E., T (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74.

The Genome of the Netherlands Consortium (2014). Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nature genetics*, 46(8):818–825.

The GTEx Consortium, Ardlie, K. G., Deluca, D. S., Segrè, A. V., Sullivan, T. J., Young, T. R., Gelfand, E. T., Trowbridge, C. A., Maller, J. B., Tukiainen, T., Lek, M., Ward, L. D., Kheradpour, P., Iriarte, B., Meng, Y., Palmer, C. D., Esko, T., Winckler, W., Hirschhorn, J. N., Kellis, M., MacArthur, D. G., Getz, G., Shabalin, A. A., Li, G., Zhou, Y.-H., Nobel, A. B., Rusyn, I., Wright, F. A., Lappalainen, T., Ferreira, P. G., Ongen, H., Rivas, M. A., Battle, A., Mostafavi, S., Monlong, J., Sammeth, M., Mele, M., Reverter, F., Goldmann, J. M., Koller, D., Guigó, R., McCarthy, M. I., Dermitzakis, E. T., Gamazon, E. R., Im, H. K., Konkashbaev, A., Nicolae, D. L., Cox, N. J., Flutre, T., Wen, X., Stephens, M., Pritchard, J. K., Tu, Z., Zhang, B., Huang, T., Long, Q., Lin, L., Yang, J., Zhu, J., Liu, J., Brown, A., Mestichelli, B., Tidwell, D., Lo, E., Salvatore, M., Shad, S., Thomas, J. A., Lonsdale, J. T., Moser, M. T., Gillard, B. M., Karasik, E., Ramsey, K., Choi, C., Foster, B. A., Syron, J., Fleming, J., Magazine, H., Hasz, R., Walters, G. D., Bridge, J. P., Miklos, M.,

Sullivan, S., Barker, L. K., Traino, H. M., Mosavel, M., Siminoff, L. A., Valley, D. R., Rohrer, D. C., Jewell, S. D., Branton, P. A., Sobin, L. H., Barcus, M., Qi, L., McLean, J., Hariharan, P., Um, K. S., Wu, S., Tabor, D., Shive, C., Smith, A. M., Buia, S. A., Undale, A. H., Robinson, K. L., Roche, N., Valentino, K. M., Britton, A., Burges, R., Bradbury, D., Hambright, K. W., Seleski, J., Korzeniewski, G. E., Erickson, K., Marcus, Y., Tejada, J., Taherian, M., Lu, C., Basile, M., Mash, D. C., Volpi, S., Struewing, J. P., Temple, G. F., Boyer, J., Colantuoni, D., Little, R., Koester, S., Carithers, L. J., Moore, H. M., Guan, P., Compton, C., Sawyer, S. J., Demchok, J. P., Vaught, J. B., Rabiner, C. A., Lockhart, N. C., Ardlie, K. G., Getz, G., Wright, F. A., Kellis, M., Volpi, S., and Dermitzakis, E. T. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235):648–660.

The International HapMap Consortium (2005). A haplotype map of the human genome. *Nature*, 437(7063):1299–320.

The International Hapmap Consortium (2010). Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52–8.

The International Hapmap Consortium, Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., Yang, H., Ch'ang, L.-Y., Huang, W., Liu, B., Shen, Y., Tam, P. K.-H., Tsui, L.-C., Waye, M. M. Y., Wong, J. T.-F., Zeng, C., Zhang, Q., Chee, M. S., Galver, L. M., Kruglyak, S., Murray, S. S., Oliphant, A. R., Montpetit, A., Hudson, T. J., Chagnon, F., Ferretti, V., Leboeuf, M., Phillips, M. S., Verner, A., Kwok, P.-Y., Duan, S., Lind, D. L., Miller, R. D., Rice, J. P., Saccone, N. L., Taillon-Miller, P., Xiao, M., Nakamura, Y., Sekine, A., Sorimachi, K., Tanaka, T., Tanaka, Y., Tsunoda, T., Yoshino, E., Bentley, D. R., Deloukas, P., Hunt, S., Powell, D., Altshuler, D., Gabriel, S. B., Zhang, H., Matsuda, I., Fukushima, Y., Macer, D. R., Suda, E., Rotimi, C. N., Adebamowo, C. A., Aniagwu, T., Marshall, P. A., Matthew, O., Nkwodimmah, C., Royal, C. D. M., Leppert, M. F., Dixon, M., Stein, L. D., Cunningham, F., Kanani, A., Thorisson, G. A., Chakravarti, A., Chen, P. E., Cutler, D. J., Kashuk, C. S., Donnelly, P., Marchini, J., McVean, G. A. T., Myers, S. R., Cardon, L. R., Abecasis, G. R., Morris, A., Weir, B. S., Mullikin, J. C., Sherry, S. T., Feolo, M., Daly, M. J., Schaffner, S. F., Qiu, R., Kent, A., Dunston, G. M., Kato, K., Niikawa, N., Knoppers, B. M., Foster, M. W., Clayton, E. W., Wang, V. O., Watkin, J., Sodergren, E., Weinstock, G. M., Wilson, R. K., Fulton, L. L., Rogers, J., Birren, B. W., Han, H., Wang, H., Godbout, M., Wallenburg, J. C., L'Archevêque, P., Bellemare, G., Todani, K., Fujita, T., Tanaka, S., Holden, A. L., Lai, E. H., Collins, F. S., Brooks, L. D., McEwen, J. E., Guyer, M. S., Jordan, E., Peterson, J. L., Spiegel, J., Sung, L. M., Zacharia, L. F., Kennedy, K., Dunn, M. G., Seabrook, R., Shillito, M., Skene, B., Stewart, J. G., Valle (chair), D. L., Clayton (co-chair), E. W., Jorde (co-chair), L. B., Cho, M. K., Duster, T., Jasperse, M., Licinio, J., Long, J. C., Ossorio, P. N., Spallone, P., Terry, S. F., Lander (chair), E. S., Lai (co-chair), E. H., Nickerson (co-chair), D. A., Boehnke, M., Douglas, J. A., Hudson, R. R., Kruglyak, L., Nussbaum, R. L., and Consortium, T. I. H. (2003). The International HapMap Project. *Nature*, 426(6968):789–796.

Therneau, T. M. (2015). *A Package for Survival Analysis in S*. version 2.38.

Thibodeau, S. N., Bren, G., and Schaid, D. (1993). Microsatellite instability in cancer of the proximal colon. *Science*, 260(May):816–819.

Thomas, H. J., Whitelaw, S. C., Cottrell, S. E., Murday, V. A., Tomlinson, I. P., Markie, D., Jones, T., Bishop, D. T., Hodgson, S. V., Sheer, D., Northover, J. M., Talbot, I. C., Solomon, E., and Bodmer, W. F. (1996). Genetic mapping of hereditary mixed polyposis syndrome to chromosome 6q. *American journal of human genetics*, 58(4):770–6.

Thompson, B. A., Spurdle, A. B., Plazzer, J.-P., Greenblatt, M. S., Akagi, K., Al-Mulla, F., Bapat, B., Bernstein, I., Capellá, G., den Dunnen, J. T., du Sart, D., Fabre, A., Farrell, M. P., Farrington, S. M., Frayling, I. M., Frebourg, T., Goldgar, D. E., Heinen, C. D., Holinski-Feder, E., Kohonen-Corish, M., Robinson, K. L., Leung, S. Y., Martins, A., Moller, P., Morak, M., Nystrom, M., Peltomaki, P., Pineda, M., Qi, M., Ramesar, R., Rasmussen, L. J., Royer-Pokora, B., Scott, R. J., Sijmons, R., Tavtigian, S. V., Tops, C. M., Weber, T., Wijnen, J., Woods, M. O., Macrae, F., Genuardi, M., and InSiGHT (2014). Application of a 5-tiered scheme for standardized classification of 2,360 unique mismatch repair gene variants in the InSiGHT locus-specific database. *Nature genetics*, 46(2):107–15.

Thompson, S. L. and Compton, D. A. (2008). Examining the link between chromosomal instability and aneuploidy in human cells. *Journal of Cell Biology*, 180(4):665–672.

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

Timpson, N. J., Walter, K., Min, J. L., Tachmazidou, I., Malerba, G., Shin, S.-Y., Chen, L., Futema, M., Southam, L., Iotchkova, V., Cocca, M., Huang, J., Memari, Y., McCarthy, S., Danecek, P., Muddyman, D., Mangino, M., Menni, C., Perry, J. R. B., Ring, S. M., Gaye, A., Dedoussis, G., Farmaki, A.-E., Burton, P., Talmud, P. J., Gambaro, G., Spector, T. D., Smith, G. D., Durbin, R., Richards, J. B., Humphries, S. E., Zeggini, E., and Soranzo, N. (2014). A rare variant in APOC3 is associated with plasma triglyceride and VLDL levels in Europeans. *Nature communications*, 5:4871.

Tiraby, J. G. and Fox, M. S. (1973). Marker discrimination in transformation and mutation of pneumococcus. *Proceedings of the National Academy of Sciences of the United States of America*, 70(12):3541–5.

Tomlinson, I. (2015). The Mendelian colorectal cancer syndromes. *Annals of clinical biochemistry*, 52(Pt 6):690–2.

Tomlinson, I., Jaeger, E., Leedham, S., and Thomas, H. (2012). Reply to "The classification of intestinal polyposis". *Nature genetics*, 45(1):2–3.

Tomlinson, I., Webb, E., Carvajal-Carmona, L., Broderick, P., Kemp, Z., Spain, S., Penegar, S., Chandler, I., Gorman, M., Wood, W., Barclay, E., Lubbe, S., Martin, L., Sellick, G., Jaeger, E., Hubner, R., Wild, R., Rowan, A., Fielding, S., Howarth, K., Silver, A., Atkin, W., Muir, K., Logan, R., Kerr, D., Johnstone, E., Sieber, O., Gray, R., Thomas, H., Peto, J., Cazier, J.-B., and Houlston, R. (2007). A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nature genetics*, 39(8):984–8.

Tomlinson, I. A. N., Rahman, N., Frayling, I. A. N., Mangion, J. O. N., Barfoot, R., Hamoudi, R., Seal, S., Northover, J., Thomas, H. U. W. J. W., Neale, K. A. Y., Hodgson, S., Talbot, I. A. N., Houlston, R., and Stratton, M. R. (1999). Inherited Susceptibility to Colorectal Adenomas and Carcinomas: Evidence for a New Predisposition Gene on 15q14-q22. *Gastroenterology*, 116(4):789–795.

Tomlinson, I. P. M., Carvajal-Carmona, L. G., Dobbins, S. E., Tenesa, A., Jones, A. M., Howarth, K., Palles, C., Broderick, P., Jaeger, E. E. M., Farrington, S., Lewis, A., Prendergast, J. G. D., Pittman, A. M., Theodoratou, E., Olver, B., Walker, M., Penegar, S., Barclay, E., Whiffin, N., Martin, L., Ballereau, S., Lloyd, A., Gorman, M., Lubbe, S., Howie, B., Marchini, J., Ruiz-Ponte, C., Fernandez-Rozadilla, C., Castells, A., Carracedo, A., Castellvi-Bel, S., Duggan, D., Conti, D., Cazier, J.-B., Campbell, H., Sieber, O., Lipton, L., Gibbs, P., Martin, N. G., Montgomery, G. W., Young, J., Baird, P. N., Gallinger, S., Newcomb, P., Hopper, J., Jenkins, M. A., Aaltonen, L. A., Kerr, D. J., Cheadle, J., Pharoah, P., Casey, G., Houlston, R. S., and Dunlop, M. G. (2011). Multiple common susceptibility variants near BMP pathway loci GREM1, BMP4, and BMP2 explain part of the missing heritability of colorectal cancer. *PLoS genetics*, 7(6):e1002105.

Tomlinson, I. P. M., Webb, E., Carvajal-Carmona, L., Broderick, P., Howarth, K., Pittman, A. M., Spain, S., Lubbe, S., Walther, A., Sullivan, K., Jaeger, E., Fielding, S., Rowan, A., Vijayakrishnan, J., Domingo, E., Chandler, I., Kemp, Z., Qureshi, M., Farrington, S. M., Tenesa, A., Prendergast, J. G. D., Barnetson, R. A., Penegar, S., Barclay, E., Wood, W., Martin, L., Gorman, M., Thomas, H., Peto, J., Bishop, D. T., Gray, R., Maher, E. R., Lucassen, A., Kerr, D., Evans, D. G. R., Schafmayer, C., Buch, S., Völzke, H., Hampe, J., Schreiber, S., John, U., Koessler, T., Pharoah, P., van Wezel, T., Morreau, H., Wijnen, J. T., Hopper, J. L., Southey, M. C., Giles, G. G., Severi, G., Castellví-Bel, S., Ruiz-Ponte, C., Carracedo, A., Castells, A., Försti, A., Hemminki, K., Vodicka, P., Naccarati, A., Lipton, L., Ho, J. W. C., Cheng, K. K., Sham, P. C., Luk, J., Agúndez, J. A. G., Ladero, J. M., de la Hoya, M., Caldés, T., Niittymäki, I., Tuupanen, S., Karhu, A., Aaltonen, L., Cazier, J.-B., Campbell, H., Dunlop, M. G., and Houlston, R. S. (2008). A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. *Nature genetics*, 40(5):623–30.

Topol, E. J., Murray, S. S., and Frazer, K. A. (2007). The genomics gold rush. *JAMA*, 298(2):218–221.

Troilo, H., Barrett, A. L., Wohl, A. P., Jowitt, T. A., Collins, R. F., Bayley, C. P., Zuk, A. V., Sengle, G., and Baldock, C. (2015). The role of chordin fragments generated by partial tolloid cleavage in regulating BMP activity. *Biochemical Society transactions*, 43(5):795–800.

Turcot, J., Despres, J. P., and St Pierre, F. (1959). Malignant tumors of the central nervous system associated with familial polyposis of the colon: report of two cases. *Diseases of the colon and rectum*, 2:465–8.

Tyrer, J., Duffy, S. W., and Cuzick, J. (2004). A breast cancer prediction model incorporating familial and personal risk factors. *Statistics in Medicine*, 23(7):1111–1130.

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., and Rozen, S. G. (2012). Primer3–new capabilities and interfaces. *Nucleic acids research*, 40(15):e115.

Valle, L., Hernández-Illán, E., Bellido, F., Aiza, G., Castillejo, A., Castillejo, M. I., Navarro, M., Seguí, N., Vargas, G., Guarinos, C., Juarez, M., Sanjuán, X., Iglesias, S., Alenda, C., Egoavil, C., Segura, Á., Juan, M. J., Rodriguez-Soler, M., Brunet, J., González, S., Jover, R., Lázaro, C., Capellá, G., Pineda, M., Soto, J. L., and Blanco, I. (2014). New insights into POLE and POLD1 germline mutations in familial colorectal cancer and polyposis. *Human Molecular Genetics*, 23(13):3506–3512.

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S., and DePristo, M. A. (2013). Current Protocols in Bioinformatics. In Bateman, A., Pearson, W. R., Stein, L. D., Stormo, G. D., and Yates, J. R., editors, *Current protocols in bioinformatics*, volume 11, pages 11.10.1–11.10.33. John Wiley & Sons, Inc., Hoboken, NJ, USA.

van Gennip, A. H., Abeling, N. G., Stroomer, A. E., van Lenthe, H., and Bakker, H. D. (1994). Clinical and biochemical findings in six patients with pyrimidine degradation defects. *Journal of inherited metabolic disease*, 17(1):130–2.

Van Gool, I. C., Eggink, F. A., Freeman-Mills, L., Stelloo, E., Marchi, E., De Bruyn, M., Palles, C., Nout, R. A., De Kroon, C. D., Osse, E. M., Klenerman, P., Creutzberg, C. L., Tomlinson, I. P. M., Smit, V. T. H. B. M., Nijman, H. W., Bosse, T., and Church, D. N. (2015). POLE proofreading mutations elicit an antitumor immune response in endometrial cancer. *Clinical Cancer Research*, 21(14):3347–3355.

Venkatachalam, R., Verwiel, E. T. P., Kamping, E. J., Hoenselaar, E., Görgens, H., Schackert, H. K., van Krieken, J. H. J. M., Ligtenberg, M. J. L., Hoogerbrugge, N., van Kessel, A. G., and Kuiper, R. P. (2011). Identification of candidate predisposing copy number variants in familial and early-onset colorectal cancer patients. *International journal of cancer*, 129(7):1635–42.

Verkasalo, P. K., Kaprio, J., Koskenvuo, M., and Pukkala, E. (1999). Genetic predisposition, environment and cancer incidence: A nationwide twin study in Finland, 1976-1995. *International Journal of Cancer*, 83(6):743–749.

Vogelstein, B., Fearon, E. R., Hamilton, S. R., Kern, S. E., Preisinger, A. C., Leppert, M., Smits, A. M. M., and Bos, J. L. (1988). Genetic Alterations during Colorectal-Tumor Development. *New England Journal of Medicine*, 319(9):525–532.

von Hansemann, D. (1890). Über asymmetrische Zellteilung in Epithelkrebsen und deren biologische Bedeutung. *Virchows Archiv für pathologische Anatomie und Physiologie und für klinische Medizin*, 119:299–326.

Vossen, C. Y., Hoffmeister, M., Chang-Claude, J. C., Rosendaal, F. R., and Brenner, H. (2011). Clotting factor gene polymorphisms and colorectal cancer risk. *J Clin Oncol*, 29(13):1722–1727.

Wada, M., Toh, S., Taniguchi, K., Nakamura, T., Uchiumi, T., Kohno, K., Yoshida, I., Kimura, A., Sakisaka, S., Adachi, Y., and Kuwano, M. (1998). Mutations in the canalicular multispecific organic anion transporter (cMOAT) gene, a novel ABC transporter, in patients with hyperbilirubinemia II/Dubin-Johnson syndrome. *Human Molecular Genetics*, 7(2):203–207.

Walter, K., Min, J. L., Huang, J., Crooks, L., Memari, Y., McCarthy, S., Perry, J. R. B., Xu, C., Futema, M., Lawson, D., Iotchkova, V., Schiffels, S., Hendricks, A. E., Danecek, P., Li, R., Floyd, J., Wain, L. V., Barroso, I., Humphries, S. E., Hurles, M. E., Zeggini, E., Barrett, J. C., Plagnol, V., Brent Richards, J., Greenwood, C. M. T., Timpson, N. J., Durbin, R., Soranzo, N., Bala, S., Clapham, P., Coates, G., Cox, T., Daly, A., Danecek, P., Du, Y., Durbin, R., Edkins, S., Ellis, P., Flicek, P., Guo, X., Guo, X., Huang, L., Jackson, D. K., Joyce, C., Keane, T., Kolb-Kokocinski, A., Langford, C., Li, Y., Liang, J., Lin, H., Liu, R., Maslen, J., McCarthy, S., Muddyman, D., Quail, M. A., Stalker, J., Sun, J., Tian, J., Wang, G., Wang, J., Wang, Y., Wong, K., Zhang, P., Barroso, I., Birney, E., Boustred, C., Chen, L., Clement, G., Cocca, M., Danecek, P., Davey Smith, G., Day, I. N. M., Day-Williams, A., Down, T., Dunham, I., Durbin, R., Evans, D. M., Gaunt, T. R., Geihs, M., Greenwood, C. M. T., Hart, D., Hendricks, A. E., Howie, B., Huang, J., Hubbard, T., Hysi, P., Iotchkova, V., Jamshidi, Y., Karczewski, K. J., Kemp, J. P., Lachance, G., Lawson, D., Lek, M., Lopes, M., MacArthur, D. G., Marchini, J., Mangino, M., Mathieson, I., McCarthy, S., Memari, Y., Metrustry, S., Min, J. L., Moayyeri, A., Muddyman, D., Northstone, K., Panoutsopoulou, K., Paternoster, L., Perry, J. R. B., Quaye, L., Brent Richards, J., Ring, S., Ritchie, G. R. S., Schiffels, S., Shihab, H. A., Shin, S.-Y., Small, K. S., Soler Artigas, M., Soranzo, N., Southam, L., Spector, T. D., St Pourcain, B., Surdulescu, G., Tachmazidou, I., Timpson, N. J., Tobin, M. D., Valdes, A. M., Visscher, P. M., Wain, L. V., Walter, K., Ward, K., Wilson, S. G., Wong, K., Yang, J., Zeggini, E., Zhang, F., Zheng, H.-F., Anney, R., Ayub, M., Barrett, J. C., Blackwood, D., Bolton, P. F., Breen, G., Collier, D. A., Craddock, N., Crooks, L., Curran, S., Curtis, D., Durbin, R., Gallagher, L., Geschwind, D., Gurling, H., Holmans, P., Lee, I., Lönnqvist, J., McCarthy, S., McGuffin, P., McIntosh, A. M., McKechanie, A. G., McQuillin, A., Morris, J., Muddyman, D., O'Donovan, M. C., Owen, M. J., Palotie, A., Parr, J. R., Paunio, T., Pietilainen, O., Rehnström, K., Sharp, S. I., Skuse, D., St Clair, D., Suvisaari, J., Walters, J. T. R., Williams, H. J., Barroso, I., Bochukova, E., Bounds, R., Dominiczak, A., Durbin, R., Farooqi, I. S., Hendricks, A. E., Keogh, J., Marenne, G., McCarthy, S., Morris, A., Muddyman, D., O'Rahilly, S., Porteous, D. J., Smith, B. H., Tachmazidou, I., Wheeler, E., Zeggini, E., Al Turki, S., Anderson, C. A., Antony, D., Barroso, I., Beales, P., Bentham, J., Bhattacharya, S., Calissano, M., Carss, K., Chatterjee, K., Cirak, S., Cosgrove, C., Durbin, R., Fitzpatrick, D. R., Floyd, J., Reghan Foley, A., Franklin, C. S., Futema, M., Grozeva, D., Humphries, S. E., Hurles, M. E., McCarthy, S., Mitchison, H. M., Muddyman, D., Muntoni, F., O'Rahilly, S., Onoufriadis, A., Parker, V., Payne, F., Plagnol, V., Lucy Raymond, F., Roberts, N., Savage, D. B., Scambler, P., Schmidts, M., Schoenmakers, N., Semple, R. K., Serra, E., Spasic-Boskovic, O., Stevens, E., van Kogelenberg, M., Vijayarangakannan, P., Walter, K., Williamson, K. A., Wilson, C., Whyte, T., Ciampi, A., Greenwood, C. M. T., Hendricks, A. E., Li, R., Metrustry, S., Oualkacha, K., Tachmazidou, I., Xu, C., Zeggini, E., Bobrow, M., Bolton, P. F., Durbin, R., Fitzpatrick, D. R., Griffin, H., Hurles, M. E., Kaye, J., Kennedy, K., Kent, A., Muddyman, D., Muntoni, F., Lucy Raymond, F., Semple, R. K., Smee, C., Spector, T. D., Timpson, N. J., Charlton, R., Ekong, R., Futema, M., Humphries, S. E., Khawaja, F., Lopes, L. R., Migone, N., Payne, S. J., Plagnol, V., Pollitt, R. C., Povey, S., Ridout, C. K., Robinson, R. L., Scott, R. H., Shaw, A., Syrris, P., Taylor, R., Vandersteen, A. M., Barrett, J. C., Barroso, I., Davey Smith, G., Durbin, R., Farooqi, I. S., Fitzpatrick, D. R., Hurles, M. E., Kaye, J., Kennedy, K., Langford, C., McCarthy, S., Muddyman, D., Owen, M. J., Palotie, A., Brent Richards, J., Soranzo, N., Spector, T. D., Stalker, J., Timpson, N. J., Zeggini, E., Amuzu, A., Pablo Casas, J., Chambers, J. C., Cocca, M., Dedoussis, G., Gambaro, G., Gasparini, P., Gaunt, T. R., Huang, J., Iotchkova, V., Isaacs, A., Johnson, J., Kleber, M. E., Kooner, J. S., Langenberg, C., Luan, J., Malerba, G., März, W., Matchan, A., Min, J. L., Morris, R., Nordestgaard, B. G., Benn, M., Ring, S., Scott, R. A., Soranzo, N., Southam, L., Timpson, N. J., Toniolo, D., Traglia, M., Tybjaerg-Hansen, A., van Duijn, C. M., van Leeuwen, E. M., Varbo, A., Whincup, P., Zaza, G., Zeggini, E., and Zhang, W. (2015). The UK10K project identifies rare variants in health and disease. *Nature*, 526(7571):82–90.

Wan, Y., Wu, W., Yin, Z., Guan, P., and Zhou, B. (2011). MDM2 SNP309, gene-gene interaction, and tumor susceptibility: an updated meta-analysis. *BMC cancer*, 11(1):208.

Wang, H., Burnett, T., Kono, S., Haiman, C. A., Iwasaki, M., Wilkens, L. R., Loo, L. W. M., Van Den Berg, D., Kolonel, L. N., Henderson, B. E., Keku, T. O., Sandler, R. S., Signorello, L. B., Blot, W. J., Newcomb, P. A., Pande, M., Amos, C. I., West, D. W., Bézieau, S., Berndt, S. I., Zanke, B. W., Hsu, L., Lindor, N. M., Haile, R. W., Hopper, J. L., Jenkins, M. A., Gallinger, S., Casey, G., Stenzel, S. L., Schumacher, F. R., Peters, U., Gruber, S. B., Tsugane, S., Stram, D. O., and Le Marchand, L. (2014a). Trans-ethnic genome-wide association study of colorectal cancer identifies a new susceptibility locus in VTI1A. *Nature communications*, 5:4613.

Wang, Y., Li, X., and Hu, H. (2014b). H3K4me2 reliably defines transcription factor binding regions in different cells. *Genomics*, 103(2-3):222–8.

Ward, L. D. and Kellis, M. (2012). HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Research*, 40(D1):D930–D934.

Warnes, G., with contributions from Gregor Gorjanc, Leisch, F., and Man., M. (2013). *genetics: Population Genetics*. R package version 1.3.8.1.

Warren, W. C., Hillier, L. W., Marshall Graves, J. A., Birney, E., Ponting, C. P., Grützner, F., Belov, K., Miller, W., Clarke, L., Chinwalla, A. T., Yang, S.-P., Heger, A., Locke, D. P., Miethke, P., Waters, P. D., Veyrunes, F., Fulton, L., Fulton, B., Graves, T., Wallis, J., Puente, X. S., López-Otín, C., Ordóñez, G. R., Eichler, E. E., Chen, L., Cheng, Z., Deakin, J. E., Alsop, A., Thompson, K., Kirby, P., Papenfuss, A. T., Wakefield, M. J., Olender, T., Lancet, D., Huttley, G. A., Smit, A. F. A., Pask, A., Temple-Smith, P., Batzer, M. A., Walker, J. A., Konkel, M. K., Harris, R. S., Whittington, C. M., Wong, E. S. W., Gemmell, N. J., Buschiazzo, E., Vargas Jentzsch, I. M., Merkel, A., Schmitz, J., Zemann, A., Churakov, G., Ole Kriegs, J., Brosius, J., Murchison, E. P., Sachidanandam, R., Smith, C., Hannon, G. J., Tsend-Ayush, E., McMillan, D., Attenborough, R., Rens, W., Ferguson-Smith, M., Lefèvre, C. M., Sharp, J. A., Nicholas, K. R., Ray, D. A., Kube, M., Reinhardt, R., Pringle, T. H., Taylor, J., Jones, R. C., Nixon, B., Dacheux, J.-L., Niwa, H., Sekita, Y., Huang, X., Stark, A., Kheradpour, P., Kellis, M., Flicek, P., Chen, Y., Webber, C., Hardison, R., Nelson, J., Hallsworth-Pepin, K., Delehaunty, K., Markovic, C., Minx, P., Feng, Y., Kremitzki, C., Mitreva, M., Glasscock, J., Wylie, T., Wohldmann, P., Thiru, P., Nhan, M. N., Pohl, C. S., Smith, S. M., Hou, S., Renfree, M. B., Mardis, E. R., and Wilson, R. K. (2008). Genome analysis of the platypus reveals unique signatures of evolution. *Nature*, 453(7192):175–183.

Warthin, A. S. (1913). Heredity with reference to carcinoma. *Arch Intern Med*, XII(5):546–555.

Wasan, H., Meade, A. M., Adams, R., Wilson, R., Pugh, C., Fisher, D., Sydes, B., Madi, A., Sizer, B., Lowdell, C., Middleton, G., Butler, R., Kaplan, R., and Maughan, T. (2014). Intermittent chemotherapy plus either intermittent or continuous cetuximab for first-line treatment of patients with KRAS wild-type advanced colorectal cancer (COIN-B): a randomised phase 2 trial. *The Lancet. Oncology*, 15(6):631–9.

Wei, W.-H., Hemani, G., and Haley, C. S. (2014). Detecting epistasis in human complex traits. *Nature Reviews Genetics*, 15(11):722–33.

Wei, Z., Wang, K., Qu, H.-Q., Zhang, H., Bradfield, J., Kim, C., Frackleton, E., Hou, C., Glessner, J. T., Chiavacci, R., Stanley, C., Monos, D., Grant, S. F. A., Polychronakos, C., and Hakonarson, H. (2009). From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes. *PLoS genetics*, 5(10):e1000678.

Welch, H. G. and Robertson, D. J. (2016). Colorectal Cancer on the Decline - Why Screening Can't Explain It All. *The New England journal of medicine*, 374(17):1605–1607.

Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–78.

Whiffin, N., Dobbins, S. E., Hosking, F. J., Palles, C., Tenesa, A., Wang, Y., Farrington, S. M., Jones, A. M., Broderick, P., Campbell, H., Newcomb, P. A., Casey, G., Conti, D. V., Schumacher, F., Gallinger, S., Lindor, N. M., Hopper, J., Jenkins, M., Dunlop, M. G., Tomlinson, I. P., and Houlston, R. S. (2013). Deciphering the genetic architecture of low-penetrance susceptibility to colorectal cancer. *Human molecular genetics*, 22(24):5075–82.

Whiffin, N., Hosking, F. J., Farrington, S. M., Palles, C., Dobbins, S. E., Zgaga, L., Lloyd, A., Kinnersley, B., Gorman, M., Tenesa, A., Broderick, P., Wang, Y., Barclay, E., Hayward, C., Martin, L., Buchanan, D. D., Win, A. K., Hopper, J., Jenkins, M., Lindor, N. M., Newcomb, P. A., Gallinger, S., Conti, D., Schumacher, F., Casey, G., Liu, T., Campbell, H., Lindblom, A., Houlston, R. S., Tomlinson, I. P., and Dunlop, M. G. (2014). Identification of susceptibility loci for colorectal cancer in a genome-wide meta-analysis. *Human molecular genetics*, 23(17):4729–37.

White, E., Patterson, R. E., Kristal, A. R., Thornquist, M., King, I., Shattuck, A. L., Evans, I., Satia-Abouta, J., Littman, A. J., and Potter, J. D. (2004). VITamins And Lifestyle cohort study: study design and characteristics of supplement users. *American journal of epidemiology*, 159(1):83–93.

Whitelaw, S., Murday, V., Tomlinson, I., Thomas, H., Cottrell, S., Ginsberg, A., Bukofzer, S., Hodgson, S., Skudowitz, R., Jass, J., Talbot, I., Northover, J., Bodmer, W., and Solomon, E. (1997). Clinical and molecular features of the hereditary mixed polyposis syndrome. *Gastroenterology*, 112(2):327–334.

Wickham, H. and Francois, R. (2015). *dplyr: A Grammar of Data Manipulation*. R package version 0.4.3.

Wigginton, J. E. and Abecasis, G. R. (2005). PEDSTATS: descriptive statistics, graphics and quality assessment for gene mapping data. *Bioinformatics*, 21(16):3445–7.

Will, O., Carvajal-Carmona, L. G., Gorman, P., Howarth, K. M., Jones, A. M., Polanco-Echeverry, G. M., Chinaleong, J. A., Günther, T., Silver, A., Clark, S. K., and Tomlinson, I. (2007). Homozygous PMS2 Deletion Causes a Severe Colorectal Cancer and Multiple Adenoma Phenotype Without Extraintestinal Cancer. *Gastroenterology*, 132(2):527–530.

Wimmer, K. and Kratz, C. P. (2010). Constitutional mismatch repair-deficiency syndrome. *Haematologica*, 95(5):699–701.

Win, A. K., Hopper, J. L., Buchanan, D. D., Young, J. P., Tenesa, A., Dowty, J. G., Giles, G. G., Goldblatt, J., Winship, I., Boussioutas, A., Young, G. P., Parry, S., Baron, J. A., Duggan, D., Gallinger, S., Newcomb, P. A., Haile, R. W., Le Marchand, L., Lindor, N. M., and Jenkins, M. A. (2013). Are the common genetic variants associated with colorectal cancer risk for DNA mismatch repair gene mutation carriers? *European journal of cancer*, 49(7):1578–87.

Win, A. K., Reece, J. C., Buchanan, D. D., Clendenning, M., Young, J. P., Cleary, S. P., Kim, H., Cotterchio, M., Dowty, J. G., MacInnis, R. J., Tucker, K. M., Winship, I. M., Macrae, F. A., Burnett, T., Le Marchand, L., Casey, G., Haile, R. W., Newcomb, P. A., Thibodeau, S. N., Lindor, N. M., Hopper, J. L., Gallinger, S., and Jenkins, M. A. (2015). Risk of colorectal cancer for people with a mutation in both a MUTYH and a DNA mismatch repair gene. *Familial Cancer*, 14(4):575–583.

Win, A. K., Walters, R. J., Buchanan, D. D., Jenkins, M. A., Sweet, K., Frankel, W. L., de la Chapelle, A., McKeone, D. M., Walsh, M. D., Clendenning, M., Pearson, S.-A., Pavluk, E., Nagler, B., Hopper, J. L., Gattas, M. R., Goldblatt, J., George, J., Suthers, G. K., Phillips, K. D., Woodall, S., Arnold, J., Tucker, K., Field, M., Greening, S., Gallinger, S., Aronson, M., Perrier, R., Woods, M. O., Green, J. S., Walker, N., Rosty, C., Parry, S., and Young, J. P. (2012). Cancer risks for relatives of patients with serrated polyposis. *The American journal of gastroenterology*, 107(5):770–8.

Wo, X., Han, D., Sun, H., Liu, Y., Meng, X., Bai, J., Chen, F., Yu, Y., Jin, Y., and Fu, S. (2011). MDM2 SNP309 contributes to tumor susceptibility: A meta-analysis. *Journal of Genetics and Genomics*, 38(8):341–350.

Wood, A. R., Perry, J. R. B., Tanaka, T., Hernandez, D. G., Zheng, H.-F., Melzer, D., Gibbs, J. R., Nalls, M. A., Weedon, M. N., Spector, T. D., Richards, J. B., Bandinelli, S., Ferrucci, L., Singleton, A. B., and Frayling, T. M. (2013). Imputation of Variants from the 1000 Genomes Project Modestly Improves Known Associations and Can Identify Low-frequency Variant - Phenotype Associations Undetected by HapMap Based Imputation. *PLoS ONE*, 8(5):e64343.

Woychik, R. P., Maas, R. L., Zeller, R., Vogt, T. F., and Leder, P. (1990). 'Formins': proteins deduced from the alternative transcripts of the limb deformity gene. *Nature*, 346(6287):850–853.

Woychik, R. P., Stewart, T. A., Davis, L. G., D'Eustachio, P., and Leder, P. (1985). An inherited limb deformity created by insertional mutagenesis in a transgenic mouse. *Nature*, 318(6041):36–40.

Wray, N. R., Goddard, M. E., and Visscher, P. M. (2007). Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Research*, 17(10):1520–1528.

Wright, S. (1926). A frequency curve adapted to variation in percentage occurrence. *Journal of the American Statistical Association*, 21(154):162–178.

Yan, K. S., Janda, C. Y., Chang, J., Zheng, G. X. Y., Larkin, K. A., Luca, V. C., Chia, L. A., Mah, A. T., Han, A., Terry, J. M., Ootani, A., Roelf, K., Lee, M., Yuan, J., Li, X., Bolen, C. R., Wilhelmy, J., Davies, P. S., Ueno, H., von Furstenberg, R. J., Belgrader, P., Ziraldo, S. B., Ordonez, H., Henning, S. J., Wong, M. H., Snyder, M. P., Weissman, I. L., Hsueh, A. J., Mikkelsen, T. S., Garcia, K. C., and Kuo, C. J. (2017). Non-equivalence of Wnt and R-spondin ligands during Lgr5(+) intestinal stem-cell self-renewal. *Nature*, 545(7653):238–242.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Goddard, M. E., and Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature genetics*, 42(7):565–9.

Ye, K., Schulz, M. H., Long, Q., Apweiler, R., and Ning, Z. (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, 25(21):2865–71.

Yeager, M., Orr, N., Hayes, R. B., Jacobs, K. B., Kraft, P., Wacholder, S., Minichiello, M. J., Fearnhead, P., Yu, K., Chatterjee, N., Wang, Z., Welch, R., Staats, B. J., Calle, E. E., Feigelson, H. S., Thun, M. J., Rodriguez, C., Albanes, D., Virtamo, J., Weinstein, S., Schumacher, F. R., Giovannucci, E., Willett, W. C., Cancel-Tassin, G., Cussenot, O., Valeri, A., Andriole, G. L., Gelmann, E. P., Tucker, M., Gerhard, D. S., Fraumeni, J. F., Hoover, R., Hunter, D. J., Chanock, S. J., and Thomas, G. (2007). Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nature genetics*, 39(5):645–649.

Zeng, C., Matsuda, K., Jia, W.-H., Chang, J., Kweon, S.-S., Xiang, Y.-B., Shin, A., Jee, S. H., Kim, D.-H., Zhang, B., Cai, Q., Guo, X., Long, J., Wang, N., Courtney, R., Pan, Z.-Z., Wu, C., Takahashi, A., Shin, M.-H., Matsuo, K., Matsuda, F., Gao, Y.-T., Oh, J. H., Kim, S., Jung, K. J., Ahn, Y.-O., Ren, Z., Li, H.-L., Wu, J., Shi, J., Wen, W., Yang, G., Li, B., Ji, B.-T., Brenner, H., Schoen, R. E., Kury, S., Gruber, S. B., Schumacher, F. R., Stenzel, S. L., Casey, G., Hopper, J. L., Jenkins, M. A., Kim, H.-R., Jeong, J.-Y., Park, J. W., Tajima, K., Cho, S.-H., Kubo, M., Shu, X.-O., Lin, D., Zeng, Y.-X., and Zheng, W. (2016). Identification of Susceptibility Loci and Genes for Colorectal Cancer Risk. *Gastroenterology*, 150(7):1633–1645.

Zhang, B., Jia, W.-H., Matsuda, K., Kweon, S.-S., Matsuo, K., Xiang, Y.-B., Shin, A., Jee, S. H., Kim, D.-H., Cai, Q., Long, J., Shi, J., Wen, W., Yang, G., Zhang, Y., Li, C., Li, B., Guo, Y., Ren, Z., Ji, B.-T., Pan, Z.-Z., Takahashi, A., Shin, M.-H., Matsuda, F., Gao, Y.-T., Oh, J. H., Kim, S., Ahn, Y.-O., Chan, A. T., Chang-Claude, J., Slattery, M. L., Gruber, S. B., Schumacher, F. R., Stenzel, S. L., Casey, G., Kim, H.-R., Jeong, J.-Y., Park, J. W., Li, H.-L., Hosono, S., Cho, S.-H., Kubo, M., Shu, X.-O., Zeng, Y.-X., and Zheng, W. (2014a). Large-scale genetic study in East Asians identifies six new loci associated with colorectal cancer risk. *Nature genetics*, 46(6):533–42.

Zhang, B., Jia, W. H., Matsuo, K., Shin, A., Xiang, Y. B., Matsuda, K., Jee, S. H., Kim, D. H., Cheah, P. Y., Ren, Z., Cai, Q., Long, J., Shi, J., Wen, W., Yang, G., Ji, B. T., Pan, Z. Z., Matsuda, F., Gao, Y. T., Oh, J. H., Ahn, Y. O., Kubo, M., Thean, L. F., Park, E. J., Li, H. L., Park, J. W., Jo, J., Jeong, J. Y., Hosono, S., Nakamura, Y., Shu, X. O., Zeng, Y. X., and Zheng, W. (2014b). Genome-wide association study identifies a new SMAD7 risk variant associated with colorectal cancer risk in East Asians. *International Journal of Cancer*, 135(4):948–955.

Zhang, J.-L., Huang, Y., Qiu, L.-Y., Nickel, J., and Sebald, W. (2007). von Willebrand factor type C domain-containing proteins regulate bone morphogenetic protein signaling through different recognition mechanisms. *The Journal of biological chemistry*, 282(27):20002–20014.

Zhou, J., Ren, K., Liu, X., Xiong, X., Hu, X., and Zhang, J. (2005). A novel PDIP1-related protein, KCTD10, that interacts with proliferating cell nuclear antigen and DNA polymerase δ. *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression*, 1729(3):200–203.

Zhu, B., Mukherjee, A., Machiela, M. J., Song, L., Hua, X., Shi, J., Garcia-Closas, M., Chanock, S. J., and Chatterjee, N. (2016). An investigation of the association of genetic susceptibility risk with somatic mutation burden in breast cancer. *British Journal of Cancer*, 115(February):1–9.

Zhu, X.-J., Liu, Y., Dai, Z.-M., Zhang, X., Yang, X., Li, Y., Qiu, M., Fu, J., Hsu, W., Chen, Y., and Zhang, Z. (2014). BMP-FGF Signaling Axis Mediates Wnt-Induced Epidermal Stratification in Developing Mammalian Skin. *PLoS Genetics*, 10(10):16.

Ziai, J., Matloff, E., Choi, J., Kombo, N., Materin, M., and Bale, A. E. (2016). Defining the polyposis/colorectal cancer phenotype associated with the Ashkenazi GREM1 duplication: counselling and management recommendations. *Genetics research*, 98:e5.

Zschocke, J., Kohlmueller, D., Quak, E., Meissner, T., Hoffmann, G. F., and Mayatepek, E. (1999). Mild trimethylaminuria caused by common variants in FMO3 gene. *Lancet*, 354(9181):834–835.

Zuk, O., Hechter, E., Sunyaev, S. R., and Lander, E. S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences of the United States of America*, 109(4):1193–8.

Zuk, O., Schaffner, S. F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., Daly, M. J., Neale, B. M., Sunyaev, S. R., and Lander, E. S. (2014). Searching for missing heritability: designing rare variant association studies. *Proceedings of the National Academy of Sciences of the United States of America*, 111(4):E455–64.

Zúñiga, A., Haramis, A. P., McMahon, A. P., and Zeller, R. (1999). Signal relay by BMP antagonism controls the SHH/FGF4 feedback loop in vertebrate limb buds. *Nature*, 401(6753):598–602.

Zúñiga, A., Laurent, F., Lopez-Rios, J., Klasen, C., Matt, N., and Zeller, R. (2012). Conserved cis-regulatory regions in a large genomic landscape control SHH and BMP-regulated Gremlin1 expression in mouse limb buds. *BMC developmental biology*, 12(1):23.

Zúñiga, A., Michos, O., Spitz, F., Haramis, A.-P. G., Panman, L., Galli, A., Vintersten, K., Klasen, C., Mansfield, W., Kuc, S., Duboule, D., Dono, R., and Zeller, R. (2004). Mouse limb deformity mutations disrupt a global control region within the large regulatory landscape required for Gremlin expression. *Genes & development*, 18(13):1553–64.