Identifying genetic biomarkers of survival for colorectal cancer

A thesis submitted in candidature for the degree of

Doctor of Philosophy (PhD)

Christopher Wills

2023

Division of Cancer and Genetics

School of Medicine

Cardiff University

Supervisors

Prof. Jeremy Cheadle

Dr Hywel Williams

Prof. Valentina Escott-Price

Contents

Abstract

Background

Clinical stage is the only routinely used marker of survival from colorectal cancer (CRC). Other factors thought to influence prognosis include the location of the primary tumour and the patient's germline and the tumour's somatic genetic profile.

Aims of my thesis

To examine inherited variation as a determinant of patient outcome with further analyses stratified by primary tumour site and mitogen-activated protein kinase (MAPK) activation status. To consider whether known somatic prognostic mutations might mask novel candidate loci.

Materials and Methods

I performed a genome-wide association study (GWAS), gene and gene-set analyses for survival in 1,926 patients with advanced CRC from the COIN and COIN-B clinical trials with replication in 5,675 patients from the Study of Colorectal Cancer in Scotland (SOCCS), 16,964 patients from the International Survival Analysis in Colorectal cancer Consortium and 5,078 patients with CRC from the UK Biobank. To understand underlying mechanism(s), I performed expression analyses both by variant and transcriptome-wide, and investigated the relationship between expression in colorectal tumours and survival in patients from The Human Protein Atlas.

Results

In COIN and COIN-B, the most significant SNP associated with survival was rs79612564 in *ERBB4* (hazard ratio [HR]=1.24, 95% confidence interval [CI]=1.16–1.32, $P=1.9\times10^{-7}$) which was replicated in stage-IV patients from SOCCS ($P=2.1\times10^{-2}$); mechanistically, patients with high *ERBB4* expression in their colon adenocarcinomas had worse survival (HR=1.50, 95% CI=1.1–1.9, $P=4.6\times10^{-2}$). When stratifying by primary tumour location, rs76011559 replicated in patients with proximal tumours (COIN, COIN-B and UK Biobank combined HR=1.53, 95% CI=1.19-1.86, $P=7.5\times10^{-7}$) and rs12273047 replicated in patients with rectal tumours (HR=1.27, 95% CI=1.09-1.46, $P=4.1\times10^{-7}$). *PI4K2B* associated with survival in patients with distal cancers ($P=2.1\times10^{-6}$) and increased *PI4K2B* expression in colorectal tumours was associated with improved survival ($P=9.6\times10^{-5}$). *RASAL2*, encoding a RAS GTPase-activating protein, was the most significant gene associated with survival in patients with MAPK-activated CRCs ($P=2.0\times10^{-5}$) with further analyses revealing pathway specificity. Finally, rs11062901 in *PARP11* was a novel biomarker of survival when unmasked from known somatic prognostic factors (HR=1.99, 95% CI=1.5-2.5, $P=4.5\times10^{-8}$) and supported by gene ($P=1.4\times10^{-6}$) and transcriptome-wide ($P=1.1\times10^{-5}$) analyses.

Conclusions

My data identify novel loci potentially associated with survival from CRC, together with mechanistic insights, many of which were mediated by changes in gene expression.

Acknowledgements

I would like to thank the following;

My lead supervisor, Prof. Jeremy Cheadle for his tireless support, guidance, and patience.

My co-supervisors, Prof. Valentina Escott-Price and Dr Hywel Williams for their feedback and guidance.

The co-authors of the publications resulting from this thesis who helped to gather and analyse data from COIN, COIN-B, SOCCS and ISACC.

Katie Watts, Amy Houseman, Dr Victoria Gray, Dr Matthew Summers, Dr Hannah West, Prof. Duncan Baird, and Miss Rachel Hargest.

Tenovus Cancer Care for funding this project.

The patients in the COIN and COIN-B clinical trials, without whom this project would not have been possible.

And finally, my family. Mum, Dad, Hannah, Matt, Nan, and Grandad. Their love, support, and hard work made me who I am today and enabled me to follow this dream.

Abbreviations

Table I. Abbreviations for amino acids

Amino acid	3-letter abbreviation	1-letter abbreviation
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamic acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

Table II. Other abbreviations

Abbreviation	Description
ADD3	*Adducin 3*
AJCC	American Joint Committee on Cancer
AMER1	*APC Membrane Recruitment Protein 1*
AMG	Amgen inc.
APC	*APC Regulator Of WNT Signalling Pathway*
ARID1A	*AT-Rich Interaction Domain 1A*
ASPECCT	A Study of Panitumumab Efficacy and Safety Compared to Cetuximab in Patients With KRAS Wild-Type Metastatic Colorectal Cancer
AUC	Area under the curve
beta	Beta-coefficient
BMP	Bone Morphogenetic Protein
BMP3	Bone Morphogenetic Protein 3
BRAF	*B-Raf Proto-Oncogene, Serine/Threonine Kinase*
CCFR	Colon Cancer Family Registry
CD/CV	Common disease, common variant
CDH1	*Cadherin 1*
CDL	Cytotoxic T lymphocytes
CEA	Carcinoembryonic antigen
CFS	Cancer-free survival
CI	Confidence interval
CIN	Chromosomal instability
COIN	COntinuous versus INtermittent
CRAN	Comprehensive R Archive Network
CRC	Colorectal cancer
CRYSTAL	Cetuximab Combined with Irinotecan in First-Line Therapy for Metastatic Colorectal Cancer
CSS	Cancer-specific survival or CRC-specific survival
CUL1	*Cullin 1*
DACHS	German Darmkrebs: Chancen der Verhutung durch Screening Study
DALS	Diet Activity and Lifestyle Study
DCC	*DCC Netrin 1 Receptor*
DFS	Disease-free survival
DNA	Deoxyribonucleic acid
DSS	Disease-specific survival
EDRN	Early Detection Research Network
EGF	Epidermal Growth Factor
EGFR	Epidermal Growth Factor Receptor

ELOVL5	*ELOVL Fatty Acid Elongase 5*
EMT	Epithelial–mesenchymal transition
EPHB1	*EPH Receptor B1*
EPIC	Swedish population of the European Prospective Investigation into Cancer
eQTL	Expression quantitative trait loci
ERBB2	*Erb-B2 Receptor Tyrosine Kinase 2*
ERK	Extracellular Signal-Regulated Kinase
et al.	*et alia* (and others)
FAP	familial adenomatous polyposis
FBXW7	*F-Box and WD Repeat Domain Containing 7*
FDR	False discovery rate
FFPE	Formalin-fixed, paraffin embedded
FFS	Failure free survival
FHIT	*Fragile Histidine Triad Diadenosine Triphosphatase*
FOLFIRI	Folinic acid, fluorouracil and irinotecan
FOLFOX	Folinic acid, fluorouracil and oxaliplatin
FPKM	Fragments per kilobase of exon per million reads
g	gram
GAP	GTPase-activating protein
GDNF	Glial Cell Line-Derived Neurotrophic Factor
GDP	Guanosine diphosphate
GNAS	*GNAS Complex Locus*
GO	Gene-ontology
GOF	Gain of function
GReX	Genetically regulated gene expression
GTEx	The Genotype-Tissue Expression project
GTP	Guanosine triphosphate
GWAS	Genome wide association study
HapMap	international haplotype map project
HCC	Hepatocellular carcinoma
HNPCC	Hereditary Non-Polyposis Colorectal Cancer/ Lynch syndrome
HPC	High performance cluster
HPFS	Health Professionals Follow-up Study
HR	Hazard ratio
HWE	Hardy-Weinberg equilibrium
IDE	Integrated development environment
IFNAR1	Interferon 1
IGF2	*Insulin Like Growth Factor 2*
INFO	Information score
IPO5	Importin 5
ISACC	International Survival Analysis in Colorectal cancer Consortium

KRAS	*Kirsten Rat Sarcoma Viral Oncogene Homolog/ KRAS Proto-Oncogene, GTPase*
LD	Linkage disequilibrium
LINC	Long Intergenic Non-Protein Coding RNA
LOF	Loss of function
LOH	Loss of heterozygosity
MAb	Monoclonal Antibody
MAD	Median absolute deviation
MAF	Minor allele frequency
MAGMA	Multi-marker Analysis of GenoMic Annotation
MAP	*MUTYH*-associated polyposis
MAPK	Mitogen-Activated Protein Kinase
MCCS	Melbourne Collaborative Cohort Study
mCRC	Metastatic colorectal cancer
miR	Micro RNA
MIR7515	*MicroRNA 7515*
MSI	Microsatellite instability
mTOR	*Mechanistic Target Of Rapamycin Kinase*
MWAS	Methylome wide association study
MYC	*MYC Proto-Oncogene, BHLH Transcription Factor*
n	Number
n/k	Not known
NDRG4	*NDRG Family Member 4*
NES	Normalised effect size
NHS	Nurses' Health Study
NRAS	*Neuroblastoma RAS Viral Oncogene Homolog/ NRAS Proto-Oncogene, GTPase*
NSAID	Non-steroidal anti-inflammatory drug
OS	overall survival
P	*P*-value
p	p-arm of a chromosome
PARP11	*Poly(ADP-Ribose) Polymerase Family Member 11*
PC	Principal component
PCA	Principal component analysis
PFS	Progression-free survival
PHS	Physicians Health Study
PI3K	phosphoinositide 3-kinase
PI4K2B	*Phosphatidylinositol 4-Kinase Type 2 Beta*
PIK3CA	*Phosphatidylinositol-4,5-Bisphosphate 3-Kinase Catalytic Subunit Alpha*
PLCO	Prostate, Lung, Colorectal, and Ovarian Study

PRIME	Panitumumab Randomized Trial In Combination With Chemotherapy for Metastatic Colorectal Cancer to Determine Efficacy
PRKCQ	*Protein Kinase C Theta*
PRKCQ-AS1	*PRKCQ Antisense RNA 1*
PTEN	*Phosphatase And Tensin Homolog*
q	q-value or q-arm of a chromosome
QC	Quality control
QQ	Quantile-quantile
RASAL2	*RAS Protein Activator Like 2*
RECIST	Response Evaluation Criteria In Solid Tumours
RET	*Ret Proto-Oncogene*
RNA	Ribonucleic acid
rsID	Unique identifier for a single nucleotide polymorphism
RSPO2	*R-Spondin 2*
RSPO3	*R-Spondin 3*
RYR3	*Ryanodine Receptor 3*
SE	Standard error
SEPT9	Septin 9
SMAD4	*SMAD Family Member 4*
SNP	Single Nucleotide Polymorphism
SOCCS	Study of Colorectal Cancer in Scotland
SOX9	*SRY-Box Transcription Factor 9*
SSM	Simple somatic mutation
TCF4	*Transcription Factor 4*
TCF7L2	*Transcription Factor 7 Like 2*
TCGA	The Cancer Genome Atlas
TGFBR2	*Transforming Growth Factor Beta Receptor 2*
TGF-α	Transforming Growth Factor Alpha
TGF-β	Transforming Growth Factor Beta
THPA	The Human Protein Atlas
TME	Tumour microenvironment
TP53	*Tumour Protein P53*
T_{reg}	Regulatory T cells
TSG	tumour-suppressor gene
TWAS	Transcriptome wide association study
U/L	Units per litre
UICC	Union for International Cancer Control
UKB	UK Biobank
VIM	*Vimentin*
VITAL	VITamins And Lifestyle Study
WBC	White blood cell
WHI	Women's Health Initiative

WHO	World Health Organisation
WT	Wild-type
XELOX	orally administered capecitabine and intravenous oxaliplatin
λ	Genomic inflation factor
	increase
	Decrease
%	Percent

List of Figures

List of tables

Publications

Publications as a direct result of the works in this thesis:

Wills, C. et al. 2021. A genome-wide search for determinants of survival in 1926 patients with advanced colorectal cancer with follow-up in over 22,000 patients. European Journal of Cancer 159, pp. 247-258. doi: 10.1016/j.ejca.2021.09.047

Wills, C. et al. 2023. Germline variation in RASAL2 may predict survival in patients with RAS-activated colorectal cancer. Genes Chromosomes & Cancer 62(6), pp. 332-341. doi: 10.1002/gcc.23133

Publications as a result of additional work I have been involved in during the course of this PhD project:

Wills, C. et al. 2023. Relationship between 233 colorectal cancer risk loci and survival in 1,926 patients with advanced disease. Accepted for publication in BJC Reports.

Watts, K. et al. 2021. Genome-wide association studies of toxicity to oxaliplatin and fluoropyrimidine chemotherapy with or without cetuximab in 1800 patients with advanced colorectal cancer. *Int J Cancer* 149(9), pp. 1713-1722. doi: 10.1002/ijc.33739

Watts, K. et al. 2022. Genetic variation in ST6GAL1 is a determinant of capecitabine and oxaliplatin induced hand-foot syndrome. *Int J Cancer*, doi: 10.1002/ijc.34046

Chapter 1: Introduction

1.1 Colorectal cancer

1.1.1 Incidence and mortality

Colorectal cancer (CRC) is cancer of the colon or rectum. It is the 4th most common cancer in the UK accounting for 11% of all new cases diagnosed every year, nearly 120 every day (years 2016-2018). CRC is most common in males (56%), people aged 75 and over (43%) and the white ethnic group (CancerResearchUK 2023). Globally, 61% of cases originate in the colon, with the remaining 39% in the rectum (Rawla et al. 2019). CRC is 3-4 times more common in developed than in developing countries, possibly due to differences in diet, physical exercise levels and ageing populations (Kuipers et al. 2015; Rawla et al. 2019).

There are approximately 16,800 CRC deaths in the UK every year, 46 every day, accounting for 10% of total cancer deaths and making CRC the 2nd biggest cancer killer (years 2017-2019). From 2009 to 2019 CRC mortality reduced by 11% in the UK (9% in females and 13% in males) and are projected to fall by an additional 10% between 2025 and 2040. The survival rate for CRC has approximately doubled in the last 40 years in the UK, with ~60% of patients surviving at least 5 years thanks to better therapeutics and public awareness (CancerResearchUK 2023). In Europe half of all cases will develop metastases, with half of those presenting with metastases at diagnosis (Haggar and Boushey 2009; Riihimäki et al. 2016). CRC can be difficult to diagnose early due to the

asymptomatic nature of the early stages of disease, initial symptoms such as blood in the stool, irregular bowel movements and weight loss can also be misdiagnosed as more common and less severe conditions.

1.1.2 CRC staging

Understanding disease stage is vital for determining prognosis and informing treatment approaches. For decades, the gold standard for tumour staging has been the American Joint Committee on Cancer (AJCC) staging manual (now in its 8th edition), which has been deployed globally by the AJCC and its partner, the Union for International Cancer Control (UICC) (Amin et al. 2017; Keung and Gershenwald 2018). This system allows solid tumours to be classified according to invasion depth (T stage), lymph node involvement (N stage) and the presence of distant metastases (M stage; **Table 1.1**). The staging system is widely accepted due to its simplicity and clinical utility due to its association with overall survival (OS) (Kattan et al. 2016). Stage IV metastatic CRC is hereby referred to as mCRC.

Stage		Tumour Size (T)	TNM Staging Lymph nodes (N)	Metastasis (M)	Description
0		Tis	N0	M0	Tumour restricted to mucosa
I		T1/T2	N0	M0	Infiltration into submucosa or muscularis propria
II	A	T3	N0	M0	Infiltration into subserosa or non-peritonealised pericolic or perirectal tissue
	B	T4a	N0	M0	Infiltration of the serosa
	C	T4b	N0	M0	Infiltration of neighbouring tissues or organs
III	A	T1-T2	N1	M0	Infiltration into submucosa or muscularis propria. Cancer cells detectable in 1-3 regional lymph nodes
		T1	N2a	M0	Infiltration into submucosa. Cancer cells detectable in 4-6 regional lymph nodes
	B	T3-T4a	N1	M0	Infiltration up to serosa. Cancer cells detectable in 1-3 regional lymph nodes
		T2-T3	N2a	M0	Infiltration into subserosa or non-peritonealised pericolic or perirectal tissue. Cancer cells detectable in 4–6 regional lymph nodes
		T1-T2	N2b	M0	Infiltration into submucosa or muscularis propria. Cancer cells detectable in 7 or more regional lymph nodes
	C	T4a	N2a	M0	Infiltration of the serosa. Cancer cells detectable in 4–6 regional lymph nodes
		T3-T4a	N2b	M0	Infiltration up to serosa. Cancer cells detectable in 7 or more regional lymph nodes
		T4b	N1-N2	M0	Infiltration of neighbouring tissues or organs. Cancer cells detectable in regional lymph nodes
IV	A	Any	Any	M1a	Metastasis to 1 distant organ or distant lymph nodes
	B	Any	Any	M1b	Metastasis to more than 1 distant organ or set of distant lymph nodes or peritoneal metastasis

Table 1.1. TNM staging of colorectal carcinoma and corresponding descriptions.

Adapted from Brenner *et al.* (2014).

1.1.3 Colorectal tumourigenesis

1.1.3.1 Risk factors

CRC is a complex disease influenced by both lifestyle and genetic factors (Kuipers et al. 2015) and unlike other common cancers no single factor accounts for the majority of cases (Brenner et al. 2014).

Studies have estimated 16-71% of CRC cases in Europe and the United States are due to lifestyle factors (Platz et al. 2000; Aleksandrova et al. 2014; Erdrich et al. 2015) which could explain the socioeconomic and geographical differences in CRC incidence (Doubeni et al. 2012). The risk of CRC increases 2-3% with each unit of body mass index (Kuipers et al. 2015) with type II diabetes patients also having an increased risk (Guraya 2015). An alcohol consumption of 2-3 units per day increases risk by 20%, with much higher consumption associated with an up to 50% increase (Fedirko et al. 2011). Prolonged heavy smoking of tobacco conveys an increase of similar magnitude (Botteri et al. 2008; Liang et al. 2009). Red and processed meat intake increases risk 16% per 100g of daily intake, whereas risk is reduced 10% per daily intake of every 10g of fibre, 200ml of milk or 300mg of calcium (Dahm et al. 2010; Song et al. 2015). Exercising for 30 minutes a day has a similar magnitude of effect (Arem et al. 2014). Use of aspirin and other NSAIDs (Algra and Rothwell 2012), statin (Bardou et al. 2010; Liu et al. 2014) and hormone therapy in postmenopausal women (Limsui et al. 2012) may also reduce risk.

1.1.3.2 CRC genetic factors

There are three common inherited CRC syndromes accounting for 2-5% (Jasperson et al. 2010) of all cases: familial adenomatous polyposis (FAP) (Fearnhead et al. 2001), *MUTYH*-associated polyposis (MAP) (Al-Tassan et al. 2002) and Hereditary Non-Polyposis Colorectal Cancer (HNPCC, also known as Lynch syndrome) (Lynch and de la Chapelle 2003; Lynch et al. 2009).

Rarer CRC syndromes include Peutz-Jeghers syndrome, an autosomal dominant disorder caused by germline mutations in the *STK-11* gene. Patients develop hamartomatous polyps of the small bowel and carry a lifetime CRC risk of 39% and near 90% for any malignancy (Kastrinos and Syngal 2011). Juvenile polyposis is another CRC syndrome characterised by multiple juvenile polyps throughout the gastrointestinal tract and a 40% lifetime risk of CRC; 40% of cases are attributed to autosomal dominant germline mutations in *SMAD4* and *BMPR1a*, with the rest not yet understood (Kastrinos and Syngal 2011). *MBD4*-associated neoplasia syndrome is an extremely rare predisposition syndrome. Like *MUTYH*, *MBD4* encodes a glycosylase of the DNA based excision repair system and germline mutations in *MBD4* have shown an autosomal recessive mode of inheritance for predisposition to CRC, acute myeloid leukaemia, gastrointestinal polyposis, uveal melanoma and schwannoma (Terradas et al. 2023). Mixed polyposis syndrome is an autosomal dominant condition characterised by an increased risk of CRC, multiple histologic polyps, including adenomas, hamartomas, and serrated lesions. Some affected individuals have been found to have germline mutations in *GREM1*, which regulates organogenesis, body patterning, and tissue differentiation,

but the genetic basis in most families is unclear (Chen et al. 2022). Polymerase proofreading-associated polyposis is an autosomal dominant adenomatous polyposis syndrome caused by germline variants in the exonuclease domains of *POLE* and *POLD1*. Although the clinical presentation remains unclear, patients exhibit a high penetrant susceptibility to CRC, polyposis and other extracolonic tumours (Chen et al. 2022). The majority of other CRC cases are sporadic and occur via the accumulation of somatic mutations and epigenetic alterations. Two distinct types of genetic mutation initiate and drive colorectal tumourigenesis; Gain of function (GOF) of oncogenes and loss of function (LOF) of tumour-suppressor genes (TSGs) (Fearon 2011).

Proto-oncogenes are a set of genes that when mutated cause normal cells to become cancerous. When mutated they are referred to as oncogenes and are most often involved in stimulating cell division, inhibiting cell differentiation, and preventing cell death, all necessary for tumour formation. The activating mutations cause the gene to either be continually transcribed or the resultant protein to be more active than its analogous wild-type. These mutations are often dominant in nature; they require only a single allele to be mutated for a cancerous phenotype (Torry and Cooper 1991; Knudson 1996; Fearon 2011) (**Table 1.2**).

Gene or biomarker	Chromosome	Function	Molecular lesion	Frequency (%)	Reference
BRAF	7	Involved in the MAPK signalling pathway	V600E-activating mutation	8–28	(Kalady et al. 2012)
ERBB2	17	Involved in the EGF–MAPK signalling pathway	Amplification	35	(Pectasides and Bass 2015)
GNAS	20	Regulates G protein signalling	Mutation	20	(Afolabi et al. 2022)
IGF2	11	Regulates the IGF signalling pathway	Copy number gain and loss of imprinting	7 (mutation); 10 (methylation)	(Kasprzak and Adamek 2019)
KRAS	12	Regulates intracellular signalling via the MAPK pathway	Activating mutations in codons 12 or 13 but rarely in codons 61, 117 and 146	40	(Allegra et al. 2009)
MYC	8	Regulates proliferation and differentiation	Amplification	2 (mutation); 10 (CNV gain)	(Strippoli et al. 2020)
NRAS	1	Regulates the MAPK pathway	Mutation in codons 12 or 13	2	(Schirripa et al. 2015)
PIK3CA	3	Regulates the PI3K–AKT pathway	Mutations in the kinase (exon 20) and helical (exon 9) domains	20	(Kato et al. 2007)
RSPO2 and *RSPO3*	8 and 6, respectively	Ligands for LGR family receptors, and activate the WNT signalling pathway	Gene fusion and translocation	10	(Sveen et al. 2020)
SOX9	17	Regulates apoptosis	Copy number gain	9 (mutation); <5 (CNV gain)	(Testa et al. 2018)
TCF7L2	10	Regulates the WNT signalling pathway	Gene fusion and translocation	10	(Wenzel et al. 2020)

Table 1.2. Proto-oncogenes involved in colorectal cancer development. Adapted from (Kuipers et al. 2015)

TSGs operate in the opposite way to oncogenes. LOF mutations, including protein truncations, insertions or deletions (indels), epigenetic silencing and missense mutations at critical residues lead to inactivation of genes responsible for DNA damage repair, cell cycle checkpoints, proliferation, cell death and cell microenvironment (Vogelstein and Kinzler 2004). Generally, these inactivating mutations are recessive and so must co-occur in both alleles (Knudson 1996; Fearon 2011). In sporadic CRC both mutations are somatic; in inherited CRC predisposition syndromes one germline mutation already exists and so only a single somatic mutation needs to occur on the second allele, this is known as the 'two-hit' hypothesis (Knudson 1996). An example of this is the germline *APC* mutation in FAP patients (Fearnhead et al. 2001) (**Table 1.3**).

1.1.3.3 Genomic instability

Among the other molecular alterations driving CRC shown in **Table 1.4**, genomic instability occurs because of mutations in proto-oncogenes and TSGs, and is a hallmark of all human cancers (Negrini et al. 2010; Sansregret et al. 2018). There are 2 major forms of genomic instability: Chromosomal instability (CIN) and microsatellite instability (MSI).

CIN occurs in approximately 70% of CRCs resulting in large structural changes and alterations in the number of chromosomes (aneuploidy). If the changes to chromosomal number or structure occur around oncogenes or TSGs the rates of mutation are increased, which can drive colorectal tumourigenesis (Hoevenaar et al. 2020) and affect tumour aggressiveness (Orsetti et al. 2014).

MSI occurs in 15% of CRCs and is characterised by hypermutation of short segments of DNA (1-6 base-pairs) repeated up to 50 times, known as microsatellites (Richard et al. 2008; Sinicrope and Sargent 2012). LOF mutations in mismatch repair genes lead to somatic changes in the microsatellites (Kawakami et al. 2015) and the MSI phenotype is a hallmark of the hereditary CRC predisposition disorder HPNCC (Lynch and de la Chapelle 2003). Strongly associated with MSI tumours is the CpG island methylator phenotype (CIMP). CIMP is characterised by aberrant methylation of promoter CpG islands resulting in epigenetic silencing of TSGs (Toyota et al. 1999; Ogino et al. 2006).

Gene or biomarker	Chromosome	Function	Molecular lesion	Frequency (%)	Reference
APC	5	Regulates the WNT signalling pathway	Inactivating mutations	40–70	(Kwong and Dove 2009)
ARID1A	1	Member of the SWI/SNF family, and regulates chromatin structure and gene transcription	Inactivating mutations	15	(Zhao et al. 2022)
DCC	18	Netrin receptor; regulates apoptosis, is deleted but not mutated in colorectal cancer, and its role in primary cancer is still unclear	Deletion or LOH	9 (mutation); 70 (LOH)	(Kudryavtseva et al. 2016)
AMER1	X	Involved in the WNT signalling pathway	Inactivating mutations	10	(Kuipers et al. 2015)
FBXW7	4	Regulates proteasome-mediated protein degradation	Inactivating mutations	20	(Li et al. 2015a)
PTEN	10	Regulates the PI3K–AKT pathway	Inactivating mutations and loss of protein (assessed by immunohistochemistry)	10 (mutation); 30 (loss of expression)	(Salvatore et al. 2019)
RET	10	Regulates the GDNF signalling pathway	Inactivating mutations and aberrant DNA methylation	7 (mutation); 60 (methylation)	(Luo et al. 2013)
SMAD4	18	Regulates the TGFβ and BMP pathways	Inactivating mutations and deletion	25	(Alhopuro et al. 2005)
TGFBR2	3	Regulates the TGFβ pathway	Inactivating mutations	20	(Tosti et al. 2022)
TP53	17	Regulates the expression of target genes involved in cell cycle progression, DNA repair and apoptosis	Inactivating mutations	50	(Liebl and Hofmann 2021)

Table 1.3. Tumour suppressor genes involved in colorectal cancer development.

Adapted from (Kuipers et al. 2015). LOH=Loss of heterozygosity.

Gene or biomarker	Chromosome	Function	Molecular lesion	Frequency (%)	Reference
Chromosome instability	-	-	Aneuploidy	70	(Pino and Chung 2010)
CpG island methylator phenotype	-	-	Methylation of >40% of loci from a selected panel of markers	15	(Toyota et al. 1999)
Microsatellite instability	-	-	Unstable microsatellite repeats in the consensus panel	15	(Sinicrope and Sargent 2012)
Mismatch-repair genes	-	Regulate DNA mismatch repair	Loss of protein (as assessed by immunohistochemistry), methylation and inactivating mutations	1–15	(Sinicrope 2010)
SEPT9	17	-	Methylation	>90	(Song and Li 2015)
VIM, *NDRG4* and *BMP3*	10, 16 and 4, respectively	-	Methylation	75	(Müller and Győrffy 2022)
18qLOH	18	-	Deletion of the long arm of chromosome 18	50	(Ogunbiyi et al. 1998)

Table 1.4. Other molecular alterations involved in colorectal cancer development.

Adapted from (Kuipers et al. 2015). LOH=Loss of heterozygosity.

1.1.3.4 Adenoma-carcinoma sequence

Mutations in specific oncogenes and tumour suppressor genes are responsible for driving the step-wise formation of a colorectal adenoma from normal epithelial tissue, and its subsequent evolution into a carcinoma, known as the adenoma-carcinoma sequence (Leslie et al. 2002). During this process there is increasing genomic instability, reducing the mutational burden of the tissue (Pino and Chung 2010) (**Figure 1.1**).

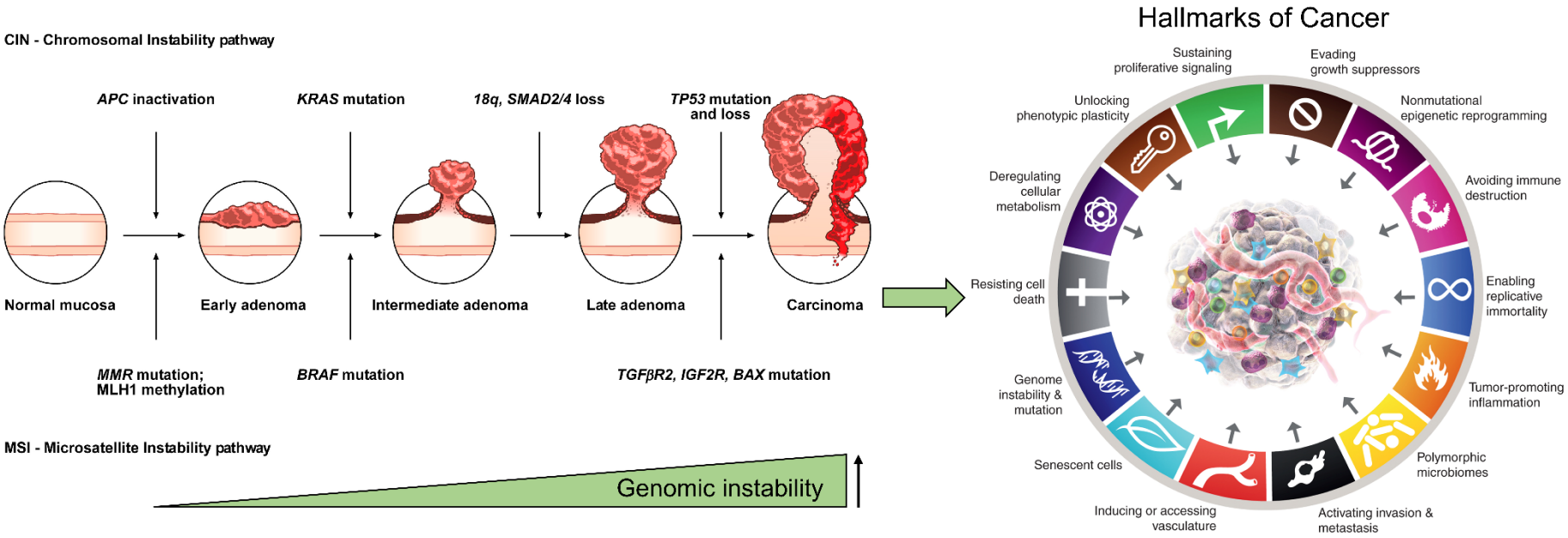


Figure 1.1. The conventional adenoma-carcinoma model of colorectal tumourigenesis. Normal mucosa form an adenoma and then a carcinoma via molecular dysregulation in one of two distinct pathways: chromosomal or microsatellite instability. The hallmarks of cancer describes the fourteen major capabilities acquired during the multistep development of cancers. Adapted from (De Palma et al. 2019) and (Hanahan 2022).

As part of the CIN pathway of CRC formation, early adenomas are formed from the biallelic inactivation of the TSG *APC*. Germline mutations in *APC* define the CRC predisposition syndrome FAP and somatic mutations occur in 40-70% of all CRCs (Muzny et al. 2012). *APC* encodes a large protein that negatively regulates the Wnt-signalling pathway. It has been associated with many commonly dysregulated processes in CRC, including cell-cycle progression, apoptosis, proliferation, polarity, stabilization of the cytoskeleton and cell-cell adhesion (Fearnhead et al. 2001).

1.1.3.5 Epidermal Growth Factor Receptor (EGFR) pathway

The EGFR signalling pathway regulates cell survival, growth, proliferation and differentiation, it is named after the transmembrane receptor for the intercellular signalling molecule epidermal growth factor (EGF) (Oda et al. 2005). EGFR, encoded by the gene Erb-B2 Receptor Tyrosine Kinase 2 (*ERBB2*), is a member of the ErbB family of receptor tyrosine kinases, is upregulated in 60-80% of CRCs and is associated with poorer prognosis (Cohen 2003).

Of the 8 EGFR ligands, EGF and transforming growth factor α (TGF-α) are the main focus of CRC research (Henriksen et al. 2013). Upon ligand binding and receptor dimerization, several signal transduction pathways are activated including the PI3K-AKT-mTOR and the MAPK/ERK (also known as RAS-RAF-MEK-ERK) pathways (**Figure 1.2**). EGFR can also be activated in a ligand-independent manner (Guo et al. 2015).

To reduce the pro-carcinogenic signalling from the binding of EGF and upregulated EGFR, several anti-EGFR therapies have been developed. Cetuximab was the first monoclonal antibody (MAb) that directly binds to the extracellular domain of EGFR inducing its internalization and degradation (Mendelsohn et al. 2015). When combined with FOLFIRI in the phase III CRYSTAL trial, there was a significant improvement in progression free survival (PFS) when compared to FOLFIRI alone (8.9 vs. 8 months, hazard ratio [HR]=0.85, *P*=0.048), however, there was no improvement in OS (HR=0.93, *P*=0.31). The apparent lack of cetuximab efficacy was later attributed to mutations in *RAS;* in the combined cetuximab treatment group samples with wild-type *RAS* showed a significant improvement in OS (HR=0.69, 95% confidence interval [CI]=0.54-0.88, $P=2.4x10^{-3}$, any *RAS* mutation HR=1.05, 95% CI=0.86-1.28, *P*=0.64) and PFS (*RAS* wild-type HR=0.56, 95% CI 0.41-0.76, *P*<0.001 any *RAS* mutation HR=1.10, 95% CI=0.85-1.42, *P*=0.47) (Van Cutsem et al. 2015). Activating mutations in *RAS* cause downstream activation of its associated pathway regardless of EGFR status, rendering EGFR inhibitors ineffective (Karapetis et al. 2008a). A 2017 meta-analysis of clinical trials involving *KRAS* wild-type mCRC patients showed that cetuximab administration was significantly associated with improved PFS (HR=0.63, 95% CI=0.50–0.79, *P*<0.0001) and OS (HR=0.74, 95% CI=0.55–0.98, *P*=0.04) (Lv et al. 2017).

Another MAb, panitumumab, also targeting EGFR was developed as an alternative to cetuximab as a fully humanized antibody which, unlike cetuximab, bares no risk of triggering antibody-dependent cell mediated cytotoxicity (Yarom and Jonker 2011). Panitumumab efficacy was assessed in the PRIME trial in a combination therapy with

FOLFOX chemotherapy; compared to FOLFOX alone the combination regimen in *KRAS* Wild-type patients showed a significant improvement in PFS (8.6 vs. 10 months, respectively; HR=0.80, 95% CI=0.66-0.97, *P*=0.02) but not OS (19.7 vs. 23.9 months; HR=0.83, 95% CI=0.67-1.02, *P*=0.072) (Douillard et al. 2010). However, OS was significant when stratified by mCRC (HR=0.83, 95% CI=0.70-0.98, *P*=0.03) (Douillard et al. 2014). No significant differences in the efficacy of cetuximab vs. panitumumab was identified in the phase III ASPECCT study (HR=0.97, *P*<0.0007 for non-inferiority) and both drugs are used as first-line mCRC treatments today.

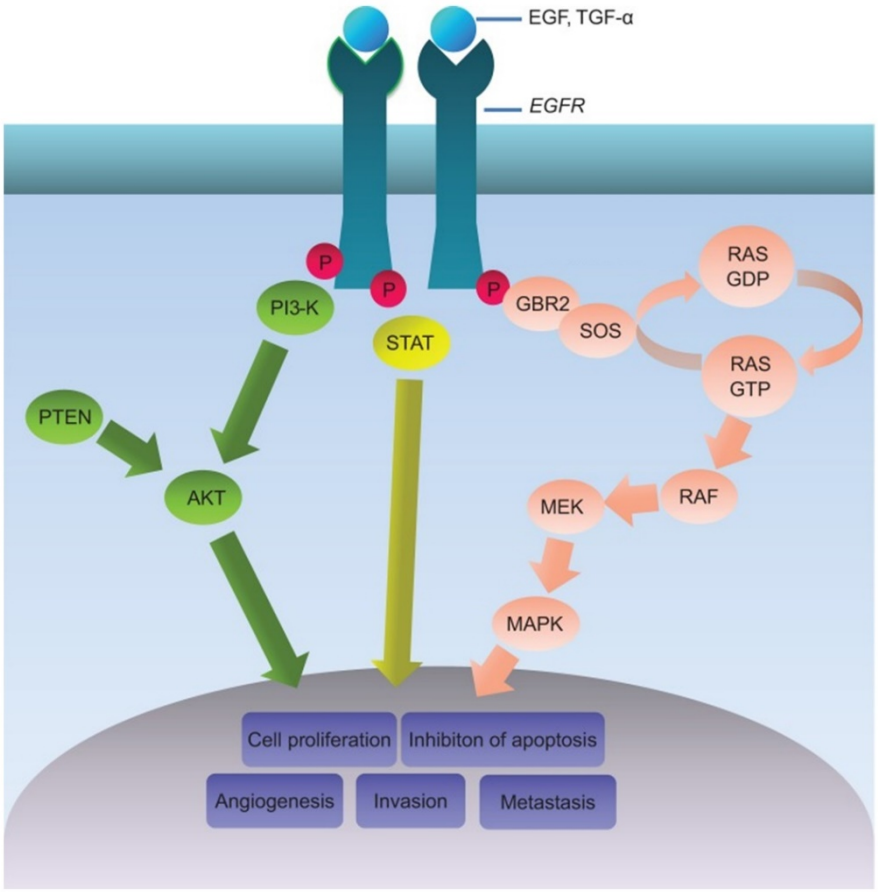


Figure 1.2. Epidermal Growth Factor Receptor (EGFR) pathway. Binding of intracellular signalling molecules such as epidermal growth factor (EGF) and transforming growth factor α (TGF-α) to EGFR triggers a cellular signalling cascade through several pathways, including the MAPK/ERK and PI3K-AKT-mTOR pathways. Resultant pro-carcinogenic behaviours include proliferation, angiogenesis, and inhibition of apoptosis. Adapted from Fang *et al.* (2014).

1.1.4 Prognostic biomarkers

1.1.4.1 Clinicopathological factors

There are many established clinicopathological factors that are predictive of CRC patient prognosis (**Table 1.5**). Females have a more favourable prognosis overall (Schmuck et al. 2020) but when analysed by age, women over 45 have a similar prognosis (i.e. statistically no significant difference) to men of the same age (Majek et al. 2013). Patients who present with a later AJCC stage at diagnosis have a significantly worse prognosis (Joachim et al. 2019). Older patients have a reduced OS (van Eeghen et al. 2015); one study showed the 5-year OS to be 0.67, 0.55 and 0.33 for patients aged <45, 45-79 and 80+ years old, respectively (McKay et al. 2014). The proximal colon (classified as the hepatic flexure, transverse colon, cecum, and ascending colon) grows from portions of the midgut and is morphologically different from both the distal (descending colon, sigmoid colon, and splenic flexure) and rectum (including the rectosigmoid junction), which grow from portions of the hindgut. Patients presenting with primary tumours in the proximal colon have a significantly worse prognosis than distal colon or rectal cancer patients (Wang et al. 2019; Bingmer et al. 2020). Patients with a greater number of metastatic sites or those whose tumours are obstructing or perforating the bowel have a worse outlook (Chen and Sheen-Chen 2000; Köhne et al. 2002). Venous invasion of cancer cells occurs in ~30% of patients, is a negative prognostic factor and can influence the decision to administer adjuvant therapies in earlier stage patients (Muller et al. 1989; Dawson et al. 2014). Several blood tests exist to screen for heightened alkaline phosphatase, platelet, and carcinoembryonic antigen levels, all of which are negative

prognostic factors (Saif et al. 2005; Stelzner et al. 2005; Wan et al. 2013).

Clinicopathological factor	Study size (n affected)	Effect on prognosis	HR	95% CI	*P*	Reference
Sex	164,996 (78,292 female)	↑ 5-year relative survival for females	-	-	<0.0001	(Majek et al. 2013)
	185,967 (85,685 female)	↑ OS for females	0.86	0.84-0.86	<0.0001	(Schmuck et al. 2020)
Stage at diagnosis	779 (486 stage III/IV)	↓ OS stage III/IV	3.70	2.89-4.99	<0.0001	(Joachim et al. 2019)
Age at diagnosis	1529 (1,459 45-79 years old)	↓ OS compared to under 45 group	1.29	0.85-1.97	<0.0001	(McKay et al. 2014)
	1529 (557 80+ years old)	↓ OS compared to under 45 group	1.95	1.27-3.01	<0.0001	
	621	↓ OS in older patients	1.02	1.01-1.04	<0.05	(van Eeghen et al. 2015)
Primary tumour location	1911 (1047 distal)	↑ OS compared to Proximal	0.72	0.62-0.83	<0.001	(Bingmer et al. 2020)
	1228 (364 rectal)	↑ OS compared to Proximal	0.75	0.61-0.92	0.006	
	1,508 (915 distal)	↑ OS compared to Proximal	0.57	0.44-0.74	<0.001	(Wang et al. 2019)
Number of metastatic sites	3825	↓ OS greater number of sites	-	-	<0.0001	(Köhne et al. 2002)
Primary tumour resection status	810 (478 resected)	↑ OS compared to unresected	0.63	0.53-0.75	<0.001	(Faron et al. 2015)
Alkaline phosphatase levels	105	↓ survival >160 U/L	4.4	1.0-19.1	-	(Saif et al. 2005)
		↓ survival >300 U/L	-	-	<0.0001	(Köhne et al. 2002)
Bowel obstruction or perforation	1837 (155 obstructed or perforated)	↓ CFS	-	-	<0.001	(Chen and Sheen-Chen 2000)
Platelet count	1,513 (231 clinically high count)	↓ OS for clinically high count	1.66	1.34-2.05	$2.6x10^{-6}$	(Wan et al. 2013)
Venous invasion	34 (6 venous invasion)	↓ survival	-	-	<0.005	(Muller et al. 1989)
WHO performance status	284 (74 performance status>2)	↓ OS	-	-	<0.001	(Strandberg Holka et al. 2018)
Carcinoembryonic antigen levels	168	↓ OS in pretherapeutic stage IV patients	2.26	1.46-3.49	0.0003	(Stelzner et al. 2005)

Table 1.5. Clinicopathological factors associated with CRC prognosis. OS=overall survival, CFS=cancer-free survival, U/L=units per litre.

1.1.4.2 Somatic mutations

Most CRC biomarker research has revolved around the acquired somatic mutations of the tumour, with many being predictive of patient survival and response to treatment (**Table 1.6**). Occurring in approximately 40% of CRCs, *KRAS* mutations are predictive of both patient prognosis (Andreyev et al. 1998; Richman et al. 2009; Eklof et al. 2013; Cremolini et al. 2015b) and response to anti-EGFR treatments (Allegra et al. 2009) due to their downstream activation of the EGFR pathway (Section 1.1.3.5). Neuroblastoma RAS Viral Oncogene Homolog (*NRAS*) mutations are a negative prognostic factor, have shown a reduction in median OS from 42.7 to 25.6 months and could also be predictive of resistance to anti-EGFR therapies (Schirripa et al. 2015). B-Raf Proto-Oncogene, Serine/Threonine Kinase (*BRAF*) mutations confer a poor prognosis (Richman et al. 2009; Kalady et al. 2012); Tran *et al.* (2011b) reported a median reduction in OS from 34.7 months to 10.4 months in *BRAF* mutants. However, approximately 90% of those *BRAF* mutation are missense mutations resulting in the V600E amino acid substitution and other $BRAF^{non\text{-}V600E}$ mutations are conversely associated with a better clinical outcome (Cremolini et al. 2015a; Schirripa et al. 2019). Phosphatidylinositol-4,5-Bisphosphate 3-Kinase Catalytic Subunit Alpha (*PIK3CA*), encoding PI3K, is a critical part of the PI3K-AKT-mTOR pathway (Section 1.1.3.5). Mutations in PIK3CA are predictive of shorter survival time (Kato et al. 2007) but is also a promising biomarker of resistance to anti-EGFR therapies due to being a downstream effector of EGFR (Cathomas 2014; Li et al. 2017). However, this treatment resistance could be restricted to exon 20 *PIK3CA* mutations (De Roock et al. 2010).

Chromosomal instability is a negative prognostic factor (Walther et al. 2008), but the prognostic role of other genomic instabilities are less clear. In contradiction to other studies (Barault et al. 2008; Kim et al. 2017), Ogino *et al.* (2009) reported a better cancer-specific survival rate for CIMP-high patients. MSI is predictive of a significantly worse survival time in mCRC patients (Tran et al. 2011b; Smith et al. 2013) but a more favourable outcome in earlier stage patients (Lochhead et al. 2013). Allelic loss at chromosome 18q, most frequently at 18q21.1, occurs in approximately 70% of CRCs and is a marker of poor prognosis (Ogunbiyi et al. 1998). Located at this locus, the TSG SMAD Family Member 4 (*SMAD4*) is commonly under expressed in CRC, resulting in a worse prognosis (Alhopuro et al. 2005).

1.1.4.3 Germline variation

Currently the only prognostic germline variant that has been robustly validated in several cohorts is the CRC-risk associated single nucleotide polymorphism (SNP; Section 1.2.1.2) rs9929218, intronic to the gene *CDH1* at 16q22.1 (**Table 1.7**). Patients homozygous for the minor A allele have a significantly worse prognosis compared to those with a copy of the major G allele, indicating a recessive model of effect (Abuli et al. 2013; Smith et al. 2015; Song et al. 2018). The variant has been shown to regulate *CDH1* expression (Han et al. 2016); *CDH1* encodes E-cadherin which controls cell polarity, adhesion, tissue morphology, cell migration and invasion of tumour cells (Takeichi 1991). Other promising prognostic germline variants include those that show primary tumour site specificity, such as rs189655236 and rs144717887 in proximal colon cancers and rs698022 in distal colon cancers (Labadie et al. 2022).

Somatic factor	Study size (n with mutation)	Effect on prognosis	HR	95% CI	*P*	Reference
KRAS mutation	689 (300)	↓ OS	1.24	1.06-1.24	$8.0x10^{-3}$	(Richman et al. 2009)
	411 (80)	↓ CSS	1.48	1.02-2.16	$2.0x10^{-3}$	(Eklof et al. 2013)
	329 (236)	↓ OS	1.49	1.11–1.99	$<1.0x10^{-4}$	(Cremolini et al. 2015b)
	2,050 (777)	↓ OS	1.22	1.07-1.40	$4.0x10^{-3}$	(Andreyev et al. 1998)
NRAS mutation	321 (47)	↓ OS	1.75	1.13-2.72	$1.3x10^{-2}$	(Schirripa et al. 2015)
BRAF mutation	692 (54)	↓ OS	1.82	1.36-2.43	$<1.0x10^{-4}$	(Richman et al. 2009)
	322 (56)	↓ OS	1.79	1.05-3.05	$3.0x10^{-2}$	(Kalady et al. 2012)
	524 (57)	↓ OS	-	-	$<1.0x10^{-3}$	(Tran et al. 2011b)
PIK3CA mutation	158 (18)	↓ DSS	-	-	$3.6x10^{-2}$	(Kato et al. 2007)
	160 (14)	↓ PFS	-	-	$3.0x10^{-2}$	(Li et al. 2017)
		↓ OS	-	-	$2.0x10^{-3}$	
MSI (mCRC)	1,565 (66)	↓ OS	1.60	1.14-2.24	$6.6x10^{-3}$	(Smith et al. 2013)
		↓ PFS	1.66	1.21-2.27	$1.6x10^{-3}$	
	350 (40)	↓ OS	-	-	$1.7x10^{-2}$	(Tran et al. 2011b)
MSI (early stages)	1,071 (92)	↑ CSS in BRAF-WT patients	0.25	0.12-0.52	$<1.0x10^{-3}$	(Lochhead et al. 2013)
CIMP-high	649 (126)	↑ colon-CSS	0.44	0.22-0.88	Significant	(Ogino et al. 2009)
	277 (37)	↓ 5-year survival in MSS patients	2.90	1.53-5.49	$<1.0x10^{-3}$	(Barault et al. 2008)
	157 (50)	↓ 5-year DFS	2.01	1.03-3.94	$4.2x10^{-2}$	(Kim et al. 2017)
CIN	10,146 (6,088)	↓ survival	1.45	1.35-1.55	$<1.0x10^{-3}$	(Walther et al. 2008)
Reduced SMAD4 Protein and mRNA levels	75 (10)	↓ DFS	-	-	Protein= $3.0x10^{-2}$	(Alhopuro et al. 2005)
			-	-	mRNA= $3.0x10^{-3}$	
Loss of hetero-zygosity at 18q	126 (67)	↓ DFS	-	-	$1.0x10^{-2}$	(Ogunbiyi et al. 1998)
		↓ DSS	-	-	$3.0x10^{-3}$	

Table 1.6. Somatic biomarkers associated with CRC prognosis. OS=Overall survival; CSS=cancer-specific survival; DSS=disease-specific survival; PFS=progression-free survival; DFS=disease-free survival; WT=wild-type

Germline SNP	Study size	Effect on prognosis	HR	95% CI	*P*	Reference
			Validated			
rs9929218	2,083	↓ OS	1.43	1.20-1.71	$5.8x10^{-5}$	(Smith et al. 2015)
	5,552	↓ OS	1.18	1.01-1.37	$3.2x10^{-2}$	(Smith et al. 2015)
	1,374	↓ OS	2.09	1.18-3.71	$1.0x10^{-2}$	(Song et al. 2018)
	1,235	↓ OS	1.54	1.06-2.22	$1.8x10^{-2}$	(Abuli et al. 2013)
			Unvalidated			
rs209489	7,258	↓ OS	1.8	1.5-2.1	$3.7x10^{-9}$	(Phipps et al. 2016)
rs10161980	5,675	↓ OS	1.24	1.10-1.39	$3.4x10^{-4}$	(He et al. 2021)
rs7495132	5,675	↓ CSS	1.97	1.41-2.74	$6.1x10^{-5}$	(He et al. 2021)
rs698022	16,964	↓ DSS	1.48	1.30-1.69	$8.47x10^{-9}$	(Labadie et al. 2022)
rs189655236	16,964	↓ DSS	2.14	1.65-2.77	$9.19x10^{-9}$	(Labadie et al. 2022)
rs144717887	16,964	↓ DSS	2.01	1.57-2.58	$3.14x10^{-8}$	(Labadie et al. 2022)

Table 1.7. Germline biomarkers associated with CRC prognosis. rs9929218, rs10161980, rs7495132 were analysed under a recessive model. rs189655236 and rs144717887 were significantly associated in proximal colon cancers and rs698022 in distal colon cancers. OS=Overall survival, CSS=CRC-specific survival, DSS=disease-specific survival.

1.2 Genome wide association studies

The genome wide association study (GWAS) is now a well-established methodology in the search for germline associations with disease phenotypes. Unlike a candidate gene study, GWAS allow for an unbiased and comprehensive scan of the whole genome (often excluding the X and Y chromosomes) without the need for prior knowledge of a particular genomic loci or biological mechanism. They allow researchers to understand complex phenotypes underlying biology, identify genetic correlations, calculate heritability, and make risk predictions. GWAS can consider sequence variations or copy-number variants but most often look for associations with SNPs (Uffelmann et al. 2021). For example, a recent GWAS meta-analysis listed 205 SNPs associated with susceptibility to CRC (n=100,204 cases and 154,587 controls of European and east Asian ancestry) (Fernandez-Rozadilla et al. 2023).

1.2.1 Underlying concepts of the GWAS design

1.2.1.1 The ‘common disease, common variant’ hypothesis

The ‘common disease, common variant’ (CD/CV) hypothesis asserts that common disorders are likely caused by genetic variants that exist in a high frequency in the population. If a common variant influences disease, then it likely has a small effect size relative to rare variants that affect rare disorders. Therefore, allele frequency and disease prevalence are inversely correlated (Manolio et al. 2009; Parikshak and Geschwind 2013). We would also expect common heritable conditions to be caused by the cumulative effect of many common variants; they are polygenic. Unrelated individuals

who are affected by a disease would share a large proportion of these low-penetrance alleles (Wang et al. 2005). In a GWAS approach to variant identification it is difficult to find rare variants with small effect sizes and these studies are often restricted to analysing common variants above a minor allele frequency (MAF) of 0.01. There are also very few examples of disease variants that are common but with high effect sizes (Manolio et al. 2009) (**Figure 1.3**).

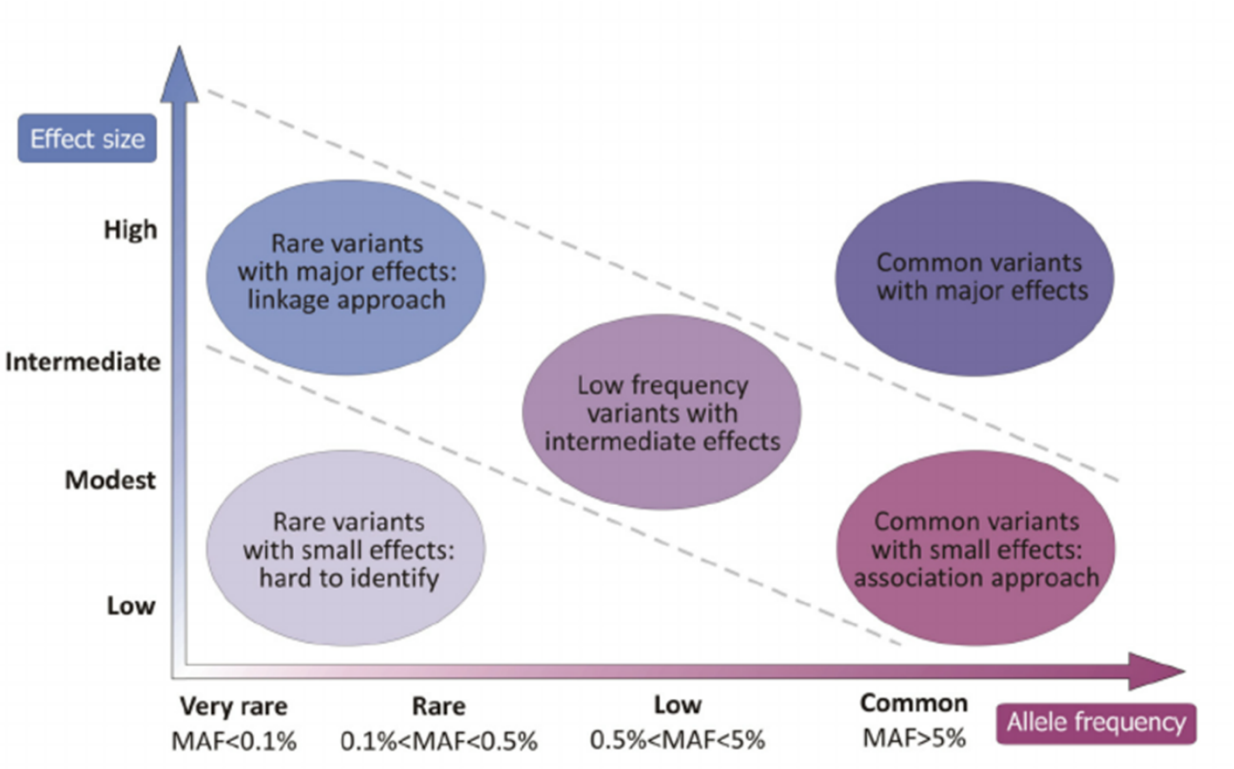


Figure 1.3. Relation of Minor Allele Frequency (MAF), effect size and feasibility of identifying risk variants by common genetic tests. The common disease, common variant hypothesis suggests that common disorders are caused by the cumulative effect of many low-penetrance variants and are studied more easily by an association analysis, such as a GWAS. Rare disorders are more likely the result of high-penetrance, rare variants identified by linkage analysis. Adapted from (Manolio et al. 2009) and (Tatijana and Vesna 2011).

1.2.1.2 Single nucleotide polymorphisms

SNPs are variations at a single position of the genome that occur in more than 1% of the population (MAF>0.01). SNPs can be a single base substitution or indel and each is assigned a unique identifier, referred to as an rsID. Approximately 90% of sequence variation in humans can be attributed to SNPs. Most SNPs are intergenic and do not impact on the structure or expression of any genes (Hunt et al. 2009). They are used in a GWAS as genetic markers of a genomic loci's association with a phenotype.

1.2.1.3 Linkage disequilibrium

During meiosis, recombination events cause exchange of genetic variants between homologous chromosomes. If two variants lie close to each other on a chromosome, then the likelihood of them being separated is reduced and they are inherited together and are in linkage disequilibrium (LD). LD is therefore a population-based parameter that describes the non-random association of two alleles (Slatkin 2008). Two measures of LD are commonly used in genetic studies, D' (used in population genetics) and r^2 (used in association studies). D' values range from -1 to 1 and are derived by dividing the coefficient of disequilibrium (D; the measure of linkage between two variants) by the theoretical maximum difference between the observed and expected allele frequencies (Lewontin 1964). r^2 values range from 0 to 1 and measure the statistical correlation between two alleles. A high r^2 value suggests that an allele for one SNP is often observed with one allele of the second SNP, meaning the two alleles are in high LD. When genotyping an individual it is therefore feasible to only genotype one of the SNPs and still

capture the allelic variation of both, allowing genotyping arrays to be a lot smaller and cheaper (Li et al. 2009).

1.2.1.4 Genotyping and imputation

GWAS most often use SNP data produced by chip-based microarrays. These arrays are cost effective and can directly genotype a few thousand to a few million SNPs. Imputation then allows for the prediction of missing SNPs, up to tens of millions, using LD information from sequenced or more densely genotyped reference populations, such as the HapMap or 1000 genomes populations (**Figure 1.4**). By imputing missing SNPs a greater coverage of the genome is achieved, increasing the statistical power and resolution to detect phenotype associations (Li et al. 2009; Howie et al. 2011). The most used imputation software is IMPUTE v2 which assigns imputed SNPs an information score between 0 and 1 indicating the likelihood that the SNP has been imputed with high certainty (Howie et al. 2009). A minimum information score threshold of 0.4 is used to filter imputed SNPs during GWAS quality control (QC) but many modern studies prefer a more stringent threshold of >0.8 to ensure the accuracy of imputed data.

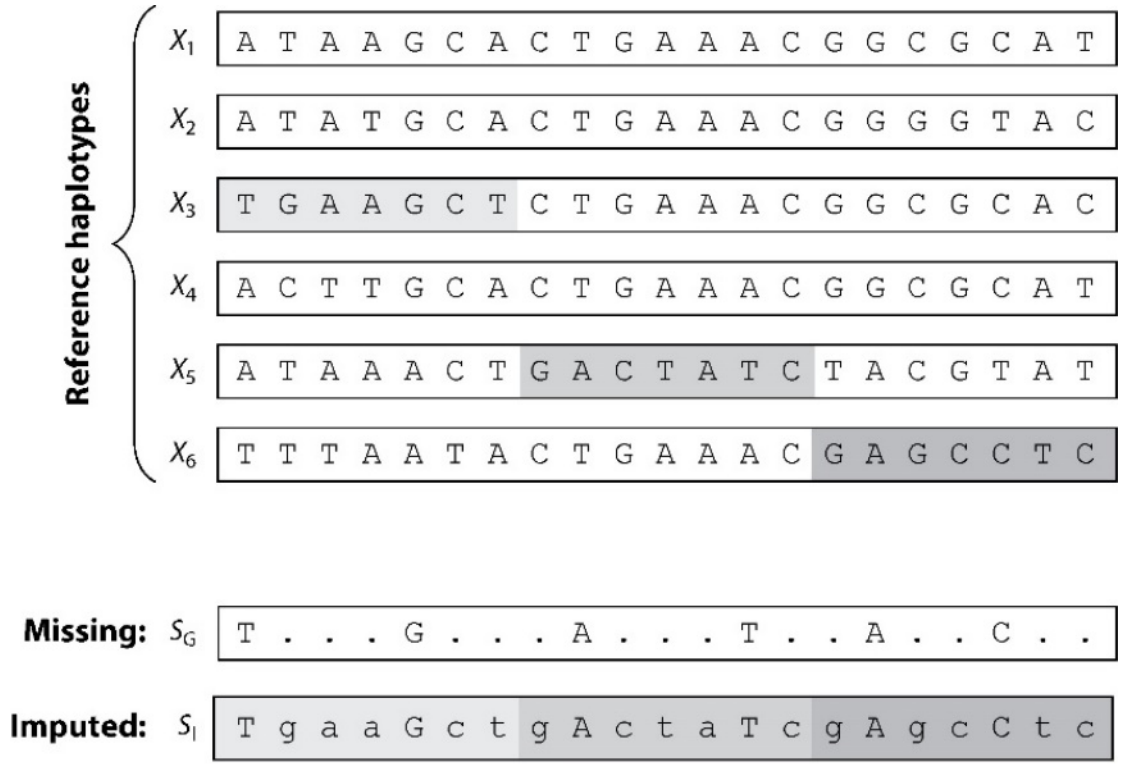


Figure 1. 4. Genotype imputation. The genotyped sample (S_G) contains untyped SNPs, using the directly genotyped SNPs it is phased with a reference population (X_n) and reference haplotypes are used to impute the untyped SNPs (S_I). Adapted from Das *et al.* (2018).

1.2.2 GWAS study design

1.2.2.1 Case-control, quantitative and time-to-event

When a trait of interest is dichotomous, a chi-squared test or logistic regression is used to compare a case group against a control group as a binary encoded phenotype. Quantitative phenotypes can also be compared under a linear model. Time-to-event phenotypes, such as survival time or time to metastatic disease, are most often analysed

using a Cox-proportional hazards model. Covariates can be added to regression models to adjust for the confounding effects of other factors, such as age and sex and reduce false-positive associations.

1.2.2.2 Genetic analysis models

Under an additive model of inheritance, each copy of a SNPs minor allele has an additive effect on the phenotype. In this case SNPs are recorded as ‘0’, ‘1’ or ‘2’ for the number of copies of the minor allele. In a recessive model, only individuals homozygous for the minor allele would have an affected phenotype and so are encoded as a ‘1’, homozygous-majors or heterozygotes are recorded as a ‘0’. Dominant alleles only require a single copy of the minor allele to have the full effect on the phenotype, this model is tested by encoding the heterozygotes and homozygous-minor samples as a 1 and the homozygous majors as a ‘0’ (Setu and Basak 2021).

1.2.2.3 Sample size, statistical power, and multiple testing

The CD/CV hypothesis proposes that common diseases are caused by SNPs with small effect sizes, as a result GWAS require very large sample sizes to be able to detect statistically significant associations. Statistical power is a measure of this ability, it is defined as the likelihood of a hypothesis test detecting a true effect if there is one and is positively linked to the sample size. It has been established that statistically significant associations from smaller, less powered studies are more likely to be false-positive findings than those identified via larger studies (Sham and Purcell 2014).

Testing millions of associations between individual SNPs and a trait of interest requires a stringent multiple testing burden to avoid false positive results. Studies such as the International HapMap Project (Altshuler et al. 2005) have shown that on average there are approximately 1 million independent common variants across the human genome, this suggests a Bonferroni corrected threshold of $P<5.0x10^{-8}$ to be suitable for GWAS and has become the de facto standard. However, when reducing the minimum MAF threshold for inclusion of rarer variants the threshold for statistical significance should be made more stringent due to the lack of LD between rare and common variants effectively increasing the number of independent tests (Uffelmann et al. 2021). More recently a second threshold for suggestive significance ($P<1.0x10^{-5}$) has been commonly accepted to identify SNPs with a potential association with the trait of interest.

1.2.3 Quality-control

1.2.3.1 Sample quality

There are stringent QC practices to remove any genetic variants or samples that may potentially bias GWAS results and lead to false-positive findings. Turner *et al.* (2011) outlined a QC protocol for GWAS data. Samples are first filtered from analysis if they contain discordant sex information (genetic sex not matching reported sex) or any large chromosomal anomalies, indicative of poor sample handling or genotyping quality. Most GWAS study designs are reliant upon the independence of the allele distributions across the study population; related samples harbour large numbers of similar genetic variants and thus bias the analyses. Commonly used tests for cryptic relatedness between samples are based on identity by descent values. In PLINK (Purcell et al. 2007) pairwise

relatedness is expressed using Pi_Hat values, a common threshold of Pi_Hat>0.1 (the minimum threshold for first cousins) is used to remove one individual from each pair. In study populations that are known to be related, genomic-relationship matrices can be calculated and incorporated in mixed model regression analyses (Widmer et al. 2014). Population stratification occurs when the study population contains different groupings of individuals of differing genetic ancestry, this can lead to the non-random assortment of alleles due to the LD structures of these sub-populations (Marchini et al. 2004). For example, if a particularly high number of individuals of a particular genetic ancestry are by chance clustered into one of the case or control groups, then all the alleles in their shared haplotype would be falsely associated with the tested phenotype. Often studies are restricted to individuals of the same genetic ancestry, identified via principal component analysis (PCA) of the genotyping data against a reference population of known ancestry, such as the 1000 genomes project (Altshuler et al. 2015). The first few genetic principal components are also often added as covariates to the regressions to further adjust for population stratification. Samples with a low genotyping call rate are also removed from analysis as this is indicative of poor-quality genotyping. The threshold used varies by study but is often >5% ungenotyped SNPs. Individuals with large deviations in genome-wide heterozygosity levels are removed; high levels indicate sample contamination and low signify inbreeding, which would bias the analysis (Marees et al. 2018a).

1.2.3.2 SNP quality

If imputed, individual SNPs are first filtered by information score (Section 1.2.1.4). SNPs that have a low call rate (more than a few percent missing) are removed, indicative of poor genotyping quality. SNPs with a MAF below 0.05 in the study population are filtered out, although many larger studies reduce this threshold to 0.01. This decreases the multiple testing burden as the power to detect an association in rare SNPs at modest effect sizes is extremely low. Rarer SNPs are also more prone to genotyping errors. Finally, variants that deviate from the Hardy-Weinberg Equilibrium (HWE) are removed as they are likely to contain genotyping errors, this is achieved using the HWE-exact test (Marees et al. 2018a).

1.2.4 GWAS visualisation

The results of the primary GWAS analysis are presented in a Manhattan plot. SNPs are ordered by chromosome then position and plotted against the association $-\log_{10}(P)$. Lines for genome wide and suggestive significance are drawn, most often at $P=5.0\times10^{-8}$ and $P=1.0\times10^{-5}$, respectively (**Figure 1.5a**). Quantile-quantile (QQ) plots are used to test for systematic inflation of *P*-values because of poor QC or model overfitting. The observed *P*-values for each SNP are ordered and plotted against expected values from a theoretical χ2-distribution (**Figure 1.5b**). If the observed values fit the expected distribution, then all points will lie along the Y=X line between the X and Y axes. Any significant SNPs observed in the study will deviate from this line but an early separation of expected from observed values may indicate QC issues, such as population stratification or cryptic relatedness (Ehret 2010). QQ plots are often accompanied by the genomic inflation factor

(λ) statistic, which is a measure of this deviation. A λ value between 1 and 1.10 is generally considered acceptable (Yang et al. 2011). Regional association plots (hereby referred to as LocusZoom plots) allow for visualisation of GWAS summary statistics at individual loci of interest. SNPs at a particular locus are plotted against their $-\log_{10}(P)$, overlapping genes are shown, as well as recombination rates and LD structure.

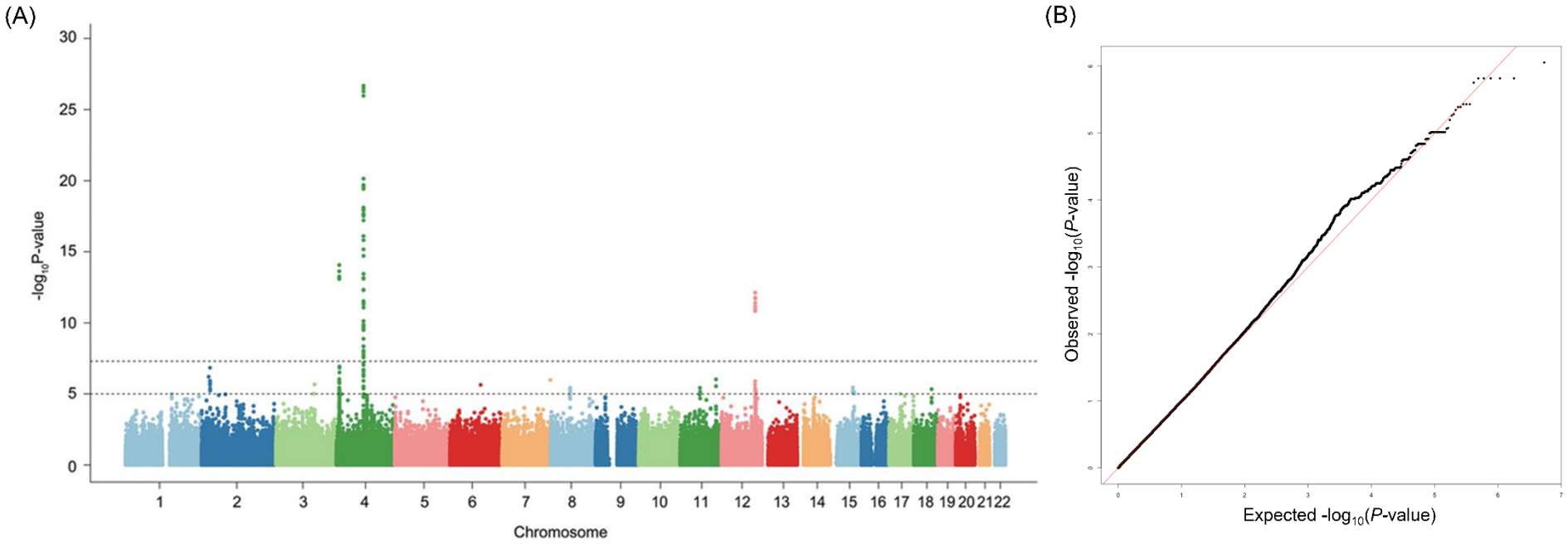


Figure 1.5. Visualisation of GWAS summary statistics. (**A**) Manhattan plot. SNPs are ordered by chromosome position and plotted against the -$\log_{10}$(*P*) for their association with gout. The top dashed line represents the threshold for genome wide significance (P=5.0x10^{-8}) and the bottom dashed line is the threshold for suggestive significance (P=1.0x10^{-5}). Adapted from (Matsuo et al. 2016) (**B**) Quantile-quantile plot: expected -$\log_{10}$(*P*-value), under the null hypothesis of no association between genotype and OS, plotted against observed -$\log_{10}$(*P*-value).

1.3 Transcriptome-wide association study

GWAS results can be difficult to interpret since strongly associated SNPs most often lie in intergenic regions of the genome and their direct effect on the phenotype of interest is unclear. These variants may regulate gene expression for nearby (cis) or more distant (trans) genes, referred to as genetically regulated gene expression (GReX). If a SNP is associated with the variance of a gene's expression (cis or trans) it is referred to as an expression quantitative trait loci (eQTL) (Nica and Dermitzakis 2013). Utilising genome-wide genotyping data and measures of gene expression (such as RNA-sequencing) there exists databases of associations between eQTLs and the tissue-specific expression of individual genes. The most commonly used databases include the Genotype-Tissue Expression (GTEx) project (Chapter 2, Section 2.3.7) and eQTLGen (Urmo et al. 2018).

Transcriptome-Wide Association Study (TWAS) is a gene-based association approach first developed by Gamazon *et al.* (2015a). Due to the limited availability of samples with directly measured transcriptome-wide gene expression levels, TWAS methods were developed to integrate genotyping or GWAS summary statistic data with reference eQTL information to identify transcriptionally regulated genes associated with a phenotype of interest. A TWAS can therefore work as an extension or alternative to a traditional GWAS approach (Li and Ritchie 2021). By aggregating the effects of many individual genetic variants into the GReX for a single gene, the multiple testing burden is reduced by orders of magnitude and significant associations are more easily interpreted as a biological mechanism of effect. TWAS approaches have previously shown success in identifying

genes whose expression is associated with CRC susceptibility (Fernandez-Rozadilla et al. 2023).

1.3.1 GWAS summary statistic-based vs individual-level data-based

TWAS first impute the transcriptome wide GReX levels using a reference panel of eQTLs and then test their association with a phenotype. What differentiates TWAS studies is the model used in the imputation of GReX levels. The two broad methods involve either the individual-level genome-wide genotyping data or summary-statistic data from a GWAS of the phenotype of interest (**Figure 1.6**). The software tool PrediXcan (Gamazon et al. 2015a) was first developed to incorporate the former but was soon followed by FUSION, developed by Gusev *et al.* (Gusev et al. 2016). FUSION was developed for use with summary-statistic level data due to the limited availability of genotyping-level data in published GWAS studies. eQTL information is highly tissue-specific and so TWAS analysis requires prior biological insight into the affected tissues of interest. More recently techniques have been developed for cross-tissue TWAS. MultiXcan by Barbeira *et al.* (2019) uses individual-level genotyping data to predict GReX in each tissue and then fits the predictions in a statistical model against the phenotype of interest. It utilises a PCA based approach to avoid inflation of results due to the correlation of cross-tissue gene expression (Li and Ritchie 2021).

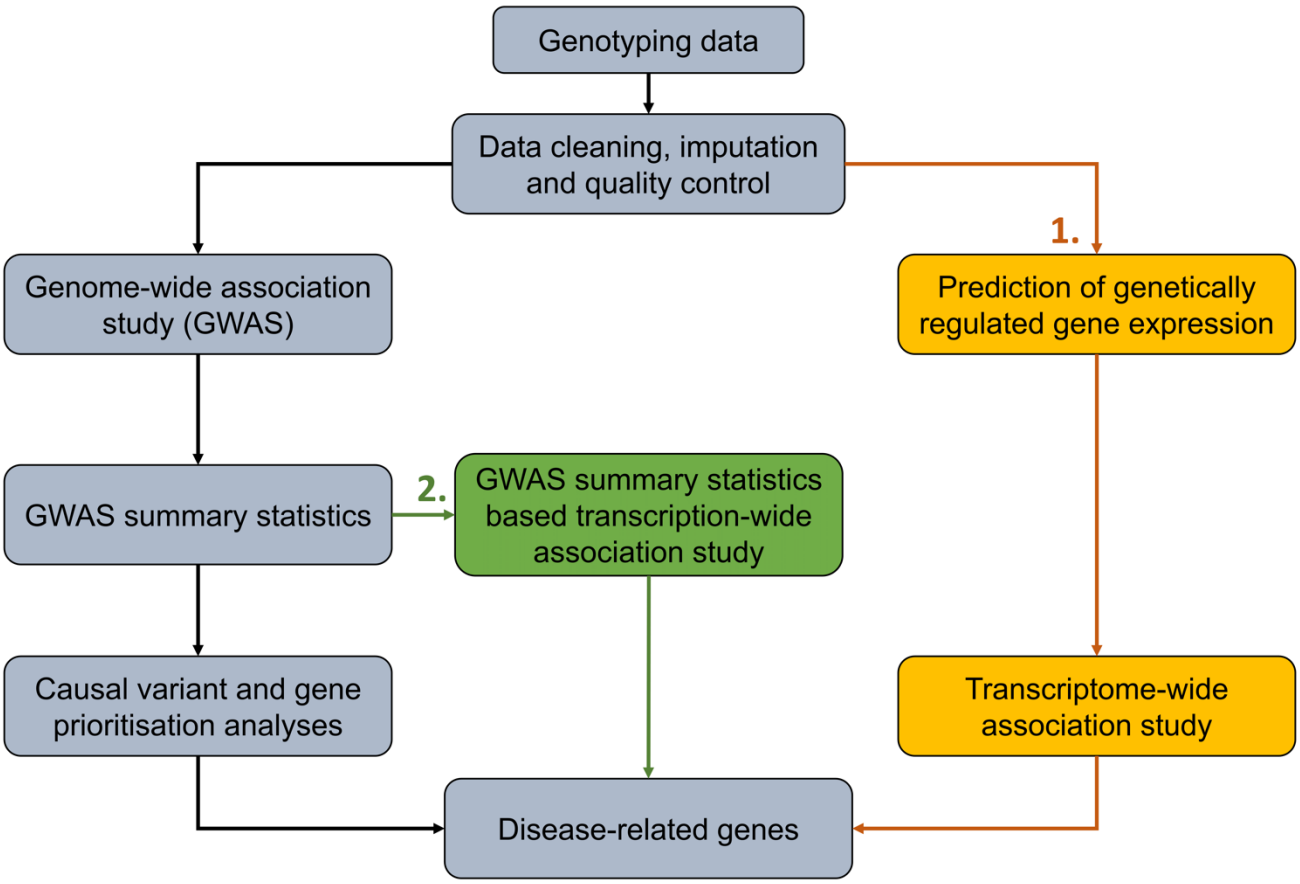


Figure 1.6. An overview of strategies for identifying disease-related genes following or parallel to GWAS. Path 1 highlights a TWAS using individual level genotyping data and path 2 a GWAS summary statistics-based TWAS. Adapted from Li and Ritchie (2021).

1.4 Hypothesis and aims

Hypothesis:

Novel germline biomarkers of survival time for CRC exist and are yet to be identified.

Aims:

- To perform a GWAS of OS in the combined COIN and COIN-B mCRC cohorts and identify novel prognostic germline alleles.
- To perform further GWAS in sub-populations grouped by primary tumour anatomical site and identify site-specific prognostic germline alleles.
- Identify potential treatment targets and germline prognostic germline alleles in MAPK-activated CRCs.
- Unmasking of novel prognostic germline alleles by excluding known somatic prognostic markers.

Figure 1.9 shows the overall structure of the thesis.

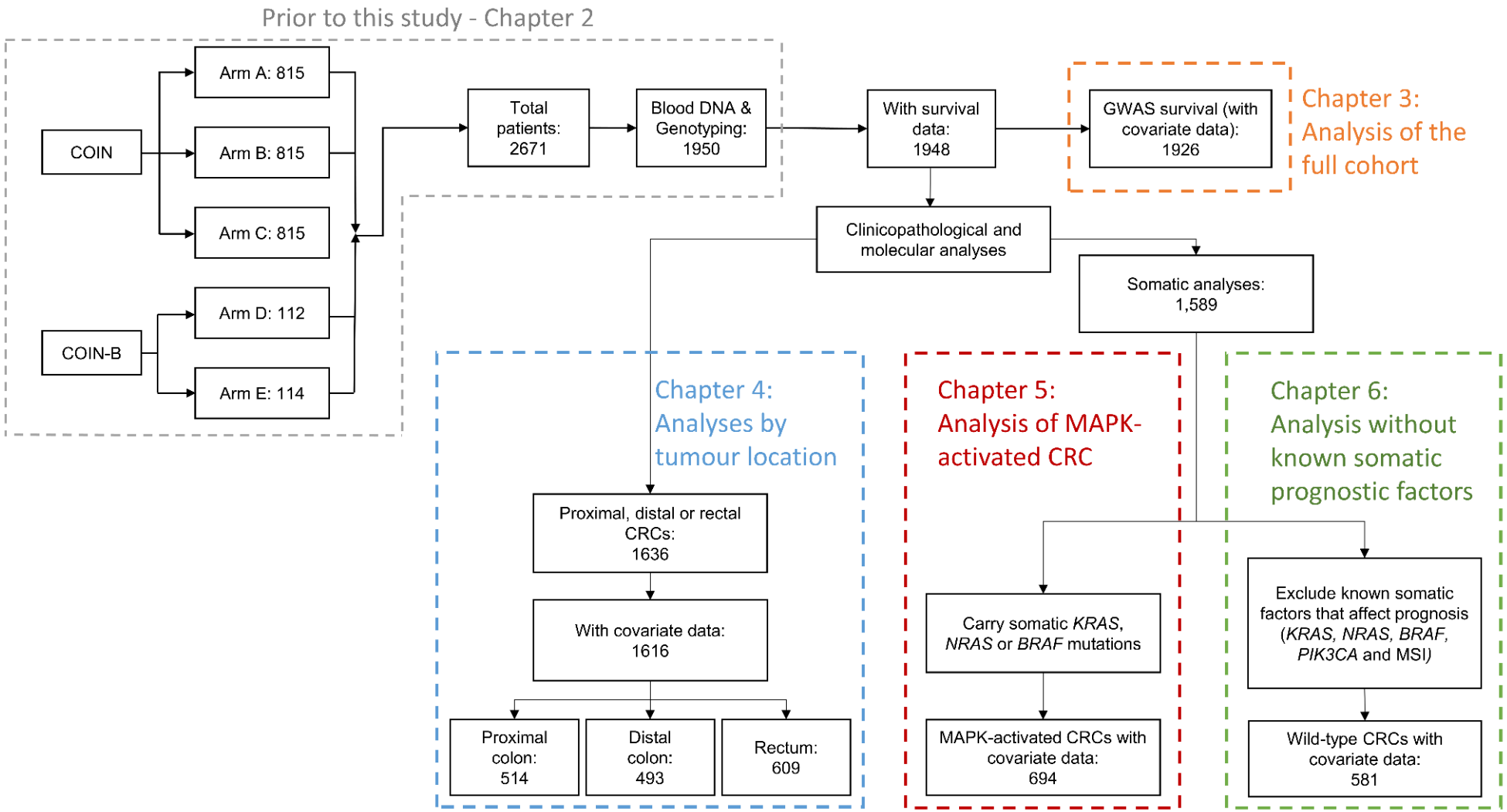


Figure 1.7. CONSORT diagram for this thesis.

Chapter 2: Materials and methods

2.1 Resources used in this thesis

2.1.1 Hardware

Local compute analyses were performed on a 2019 Apple (Cupertino, USA) MacBook Pro Retina (15", 2.4GHz 8-core Intel Core i9 processor, 32GB 24000 MHz DDR4 memory) using the macOS Monterey operating system. Analyses requiring advanced compute were completed via command line-based remote access of the Hawk high-performance cluster (HPC) located at the Advanced Research Computing at Cardiff (ARCCA) facility.

2.1.2 Software

The statistical programming language R, version 4.1.1 (R_Core_Team 2018), downloaded from http://www.r-project.org, was used for data processing and analysis. The general-purpose language Python version 3.10 was used also for data manipulation (Van Rossum and Drake 2009). The integrated development environments (IDE) used were RStudio version 2022.02.3+492 (Orange Blossom release, RStudio, Inc., Boston, MA) downloaded from https://www.rstudio.com/ and Visual Studio Code version 1.67 (Microsoft, Redmond, WA), downloaded from https://code.visualstudio.com/. Linear and logistic GWAS analyses, LD-based SNP clumping and management of the binary genotyping files were completed using PLINK versions 1.9 (Purcell et al. 2007) and 2.0 (Chang et al. 2015), downloaded from http://pngu.mgh.harvard.edu/purcell/plink/. Gene and gene-set level association

analyses were completed using Multi-marker Analysis of GenoMic Annotation (MAGMA) (de Leeuw et al. 2015) versions 1.07b (Chapter 3) and 1.09b (Chapters 4, 5 and 6), downloaded from https://ctg.cncr.nl/software/magma. SNPTEST version 2 (Marchini and Howie, Oxford, UK) was used to calculate SNP INFO scores, downloaded from https://www.well.ox.ac.uk/~gav/snptest/. GTOOL (Genomics Software Suite, University of Oxford) was used to convert genotype files, downloaded from https://www.well.ox.ac.uk/~cfreeman/software/gwas/gtool.html. The UK Biobank phenotypic and clinical dataset was decompressed and converted into tab delimited text files using the ukbunpack and ukbconv software. The genotypic data was downloaded using the gfetch software, all available from the UK Biobank website https://biobank.ndph.ox.ac.uk/showcase/download.cgi. PrediXcan (Gamazon et al. 2015b), part of the MetaXcan tool set (downloaded from https://github.com/hakyimlab/MetaXcan) was used to impute individual-level gene expression levels from genotype data.

2.1.3 Packages and Modules

Packages for R were downloaded from the Comprehensive R Archive Network (CRAN, https://cran.r-project.org/), Bioconductor (https://www.bioconductor.org/) repositories and individual Git (https://github.com/) repositories. Python modules were downloaded from the conda package management system (Anaconda_inc. 2020). All modules and packages used for this thesis are listed in **Table 2.1**.

Package/module	Software	Purpose	Reference
Base	R	Basic data manipulation	R Core Team (2018)
BiocManager	R	Used to access the Bioconductor repository of packages	Gentleman et al. (2004)
car	R	Function for recoding of variables	Fox and Sanford (2019)
data.table	R	Data import and export	Dowle and Srinivasan (2019)
gwasurvivr	R	Genome wide association analysis of time-to-event variables	Rizvi et al. (2019)
NumPy	Python	Mathematical functions	Harris et al. (2019)
Pandas	Python	Data manipulation and analysis	McKinney (2010)
Psych	R	Functions for Principal component analysis	Revelle (2021)
qqman	R	Generating Quantile-Quantile and Manhattan plots	Turner (2018)
qvalue	R	Functions to adjust *P*-values for false discovery rate	Storey et al. (2021)
survival	R	Functions for time-to-event data analysis	Therneau (2022)
survminer	R	Functions for time-to-event data visualisation	Kassambara (2021)
survSNP	R	Power calculations for SNP association studies with time-to-event data	Owzar (2012)
tidyverse	R	Collection of packages designed for data science, including ggplot2, dplyr and tibble.	Wickham et al. (2019)

Table 2.1. Packages and modules used in this thesis.

2.1.4 Web Links

Web based packages used for further analyses included LocusZoom (Willer et al. 2010a), for visualisation of GWAS summary statistics and SNP LD information, available at http://locuszoom.org.

2.2 My contribution and others contributions

Sample collection, genotyping and some QC measures were completed by others prior to the beginning of this project, all other analyses and the study design were completed by myself unless stated otherwise (**Figure 1.7**).

2.3 Datasets used in this thesis

2.3.1 COIN and COIN-B

2.3.1.1 COIN

The COIN trial (NCT00182715) was a phase III randomised clinical trial in mCRC patients for the anti-cancer drug cetuximab, a monoclonal antibody targeting EGFR (Chapter 1, Section 1.1.3.5) (Adams et al. 2011; Maughan et al. 2011). Two thousand, four hundred and forty-five patients with locally advanced or metastatic colorectal adenocarcinoma were randomised 1:1:1 into three arms. Arm A (n=815) received continuous chemotherapy (intravenous 5-FU, folinic acid (leucovorin) and oxaliplatin (FOLFOX) or orally administered capecitabine and intravenous oxaliplatin (XELOX)), Arm B (n=815) received continuous chemotherapy plus continuous cetuximab and Arm C (n=815) received intermittent chemotherapy (**Figure 2.1)**. Oxaliplatin plus

fluorouracil and folinic acid was given as a 2-weekly regimen of intravenous L-folinic acid 175 mg or D,L-folinic acid 350 mg over 2h given concurrently with oxaliplatin 85 mg/m² over 2h, followed by intravenous bolus fluorouracil 400 mg/m², and finally fluorouracil 2400 mg/m² infusion over 46h via an ambulatory pump. Oxaliplatin plus capecitabine was given as a 3-weekly regimen of intravenous oxaliplatin 130 mg/m² over 2 h followed by oral capecitabine 1000 mg/m² twice a day for 2 weeks (Adams et al. 2011). Inclusion criteria comprised of patients being at least 18 years old, primary adenocarcinoma of the colon or rectum, inoperable metastatic or locoregional measurable disease according to Response Evaluation Criteria In Solid Tumours (RECIST, version 1.0), good end-organ function and World Health Organisation (WHO) performance status of maximum 2. Patients were excluded if they had a history of malignant disease, an uncontrolled medical comorbidity likely to interfere with the trial, previous chemotherapy treatment or metastases in the brain. Patients gave informed consent for bowel cancer research (approved by REC [04/MRE06/60]).

The aims of the COIN study were to (I) assess the effect on OS of the addition of cetuximab to first-line continuous chemotherapy and (II) determine if intermittent chemotherapy was inferior to continuous chemotherapy in terms of OS. In terms of OS or PFS, there was no statistically significant superiority of cetuximab addition to continuous chemotherapy versus continuous chemotherapy alone (**Figure 2.2**), even in patients with *KRAS* wild-type CRC (OS HR=1.04, 95% CI=0.87-1.23, *P*=0.67; PFS HR=0.96, 95% CI=0.82-1.12, *P*=0.60) (Maughan et al. 2011). Intermittent chemotherapy did not show non-inferiority to continuous chemotherapy in terms of OS (median OS 19.6 months Arm A, 18.0 months Arm C; HR=1.05, 95% CI=0.85-1.29, *P*=0.66). However, subgroup analyses did show that patients with normal baseline

platelet counts could have intermittent chemotherapy, and all its associated benefits, with no detriment in survival. Patients with raised platelet counts require continuous chemotherapy to increase their survival time and quality of life (Adams et al. 2011).

2.3.1.2 COIN-B

The follow up phase II COIN-B clinical trial (NCT00640081) recruited a further 226 patients, with the same inclusion/exclusion criteria as COIN to determine the efficacy of intermittent cetuximab against cetuximab maintenance. Following the emergence of data showing the resistance of *KRAS*-mutant tumours to anti-EGFR therapies (Chapter 1, Section 1.1.3.5) trial recruitment was suspended in May 2008 and recommenced in January 2009 recruiting only *KRAS* wild-type patients. Arm D (n=112) received intermittent FOLFOX chemotherapy plus intermittent cetuximab and Arm E (n=114) received intermittent FOLFOX chemotherapy plus continuous cetuximab (Wasan et al. 2014) (**Figure 2.1**). In the analysis of 169 *KRAS* wild-type patients, continuous cetuximab showed superiority to intermittent treatment in terms of PFS (median PFS intermittent cetuximab 3.1 months, 95% CI=2.8-4.7; continuous cetuximab 5.8 months, 95% CI=4.9-8.6) and failure-free survival (FFS) (FFS intermittent cetuximab 16.8 months, 95% CI=14.5-22.6; continuous cetuximab 22.2 months, 95% CI=18.4-28.9). Clinicopathological data of patients by trial arm can be seen in **Table 2.2**.

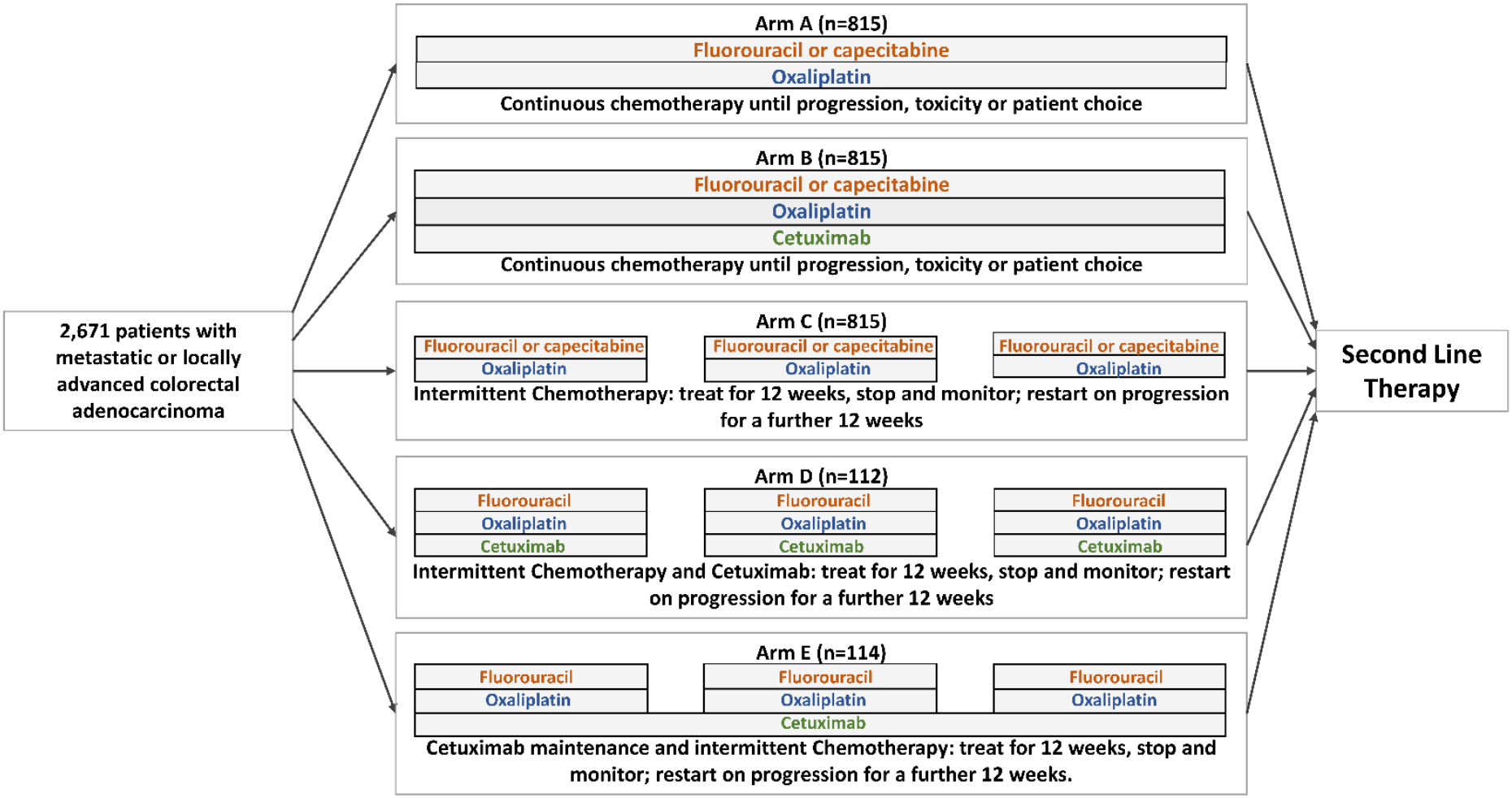


Figure 2.1. COIN and COIN-B trial design

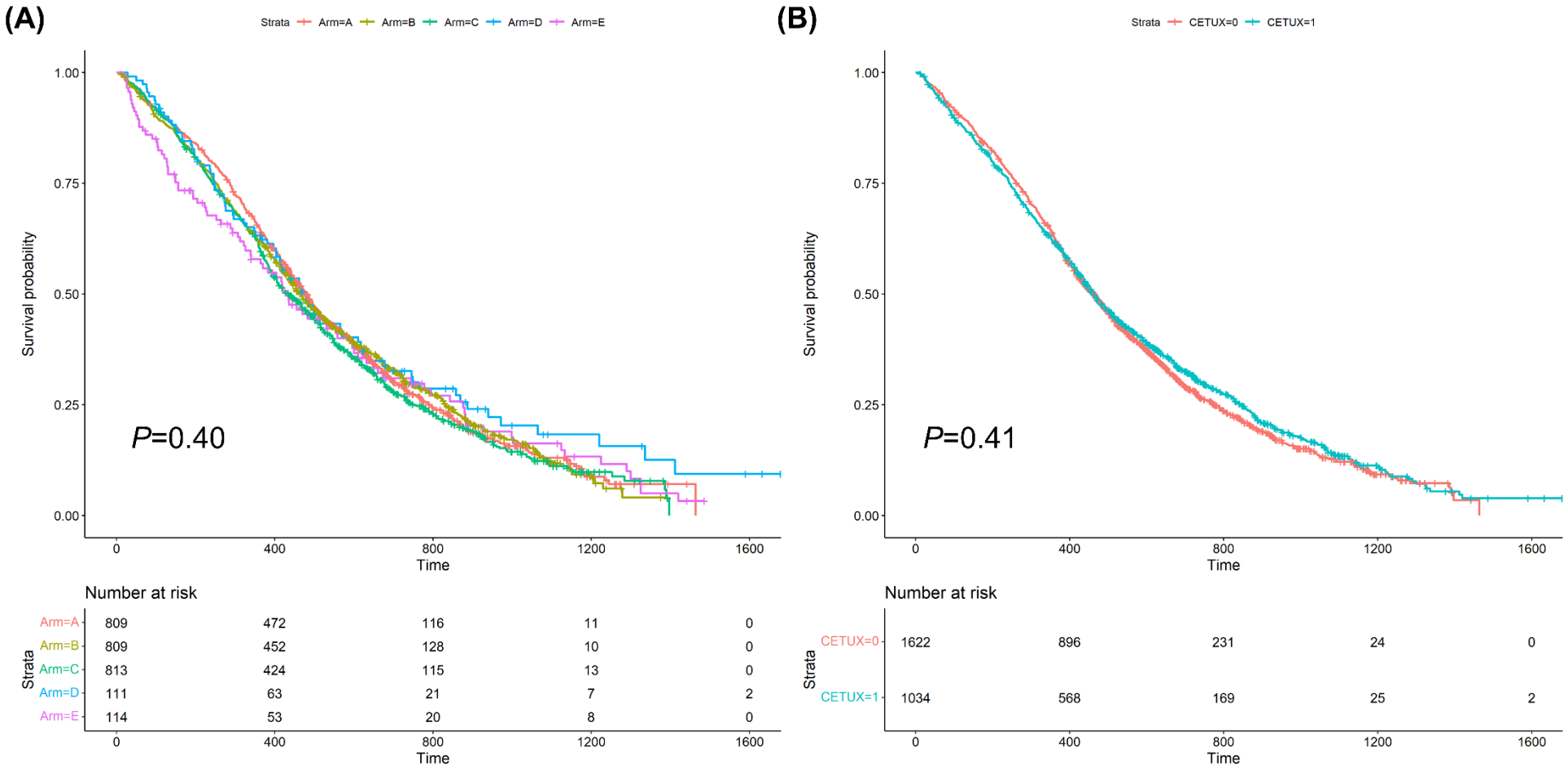


Figure 2.2. Kaplan-Meier survival analyses from the COIN and COIN-B trials. Time in days is plotted against overall survival probability for (A) patients from trial arms A-E and (B) patients who did and did not receive cetuximab. The number of patients still at risk at each time point is shown beneath and *P*-values are shown for log-rank tests.

Trial and arm		COIN A	COIN B	COIN C	COIN-B D	COIN-B E
Patients	Total	815	815	815	112	114
	Genotype and passed QC	579 (71)	616 (76)	583 (72)	85 (76)	85 (75)
Sex	Male	390 (67)	410 (67)	376 (64)	48 (56)	46 (54)
	Female	189 (33)	206 (33)	207 (36)	37 (44)	39 (46)
Mean Age		62.3	62.9	63.2	61.8	61.9
Chemotherapy received	FOLFOX	200 (35)	212 (34)	212 (36)	85 (100)	85 (100)
	XELOX	379 (65)	404 (66)	371 (64)	0 (0)	0 (0)
Cetuximab administered	Yes	0 (0)	616 (100)	0 (0)	85 (100)	85 (100)
	No	579 (100)	0 (0)	583 (100)	0 (0)	0 (0)
Primary tumour location	Colon	390 (67)	408 (66)	405 (70)	52 (61)	69 (81)
	Rectum	187 (32)	208 (34)	177 (30)	33 (39)	16 (19)
	n/k	2 (<1)	0 (0)	1 (<1)	0 (0)	0 (0)
Number of metastatic sites	0–1	197 (34)	239 (39)	208 (36)	30 (35)	32 (38)
	≥ 2	382 (66)	377 (61)	375 (64)	55 (65)	53 (62)
Liver-only metastases	Yes	432 (75)	462 (75)	440 (75)	0 (0)	0 (0)
	No	147 (25)	154 (25)	143 (25)	0 (0)	0 (0)
	n/k	0 (0)	0 (0)	0 (0)	85 (100)	85 (100)
Synchronous metastases	Yes	393 (68)	426 (69)	411 (70)	61 (72)	67 (79)
	No	180 (31)	187 (30)	167 (29)	23 (27)	18 (21)
	n/k	6 (1)	3 (<1)	5 (1)	1 (1)	0 (0)
WHO performance status	0-1	537 (93)	575 (93)	535 (92)	80 (94)	76 (89)
	≥ 2	42 (7)	41 (7)	48 (8)	5 (6)	9 (11)
White blood cell count	<10000 (per L)	404 (70)	442 (72)	399 (68)	73 (86)	63 (74)
	≥ 10000 (per L)	175 (30)	174 (28)	183 (31)	12 (14)	21 (25)
	n/k	0 (0)	0 (0)	1 (<1)	0 (0)	1 (1)
Response at 12 weeks	Yes	277 (48)	300 (49)	289 (46)	49 (58)	39 (46)
	No	218 (38)	223 (36)	210 (36)	21 (25)	23 (27)
	no data	84 (14)	93 (15)	84 (14)	15 (17)	23 (27)
Median OS (days)		503	496	461	509	527
KRAS status	Mutant	268 (33)	297 (36)	259 (32)	24 (21)	15 (13)
	Wild-type	367 (45)	362 (44)	396 (49)	78 (70)	91 (80)
	no data	180 (22)	156 (19)	160 (20)	10 (9)	8 (7)
NRAS status	Mutant	18 (2)	32 (4)	19 (2)	7 (6)	8 (7)
	Wild-type	613 (75)	627 (77)	630 (77)	62 (55)	76 (67)
	no data	184 (23)	156 (19)	166 (20)	43 (38)	30 (26)

BRAF status	Mutant	44 (8)	29 (5)	52 (9)	6 (7)	12 (14)
	Wild-type	426 (74)	480 (78)	435 (75)	46 (54)	51 (60)
	no data	109 (19)	107 (17)	96 (16)	33 (39)	22 (26)
PIK3CA status	Mutant	58 (10)	67 (11)	64 (11)	0 (0)	0 (0)
	Wild-type	400 (69)	432 (70)	419 (72)	2 (2)	0 (0)
	no data	121 (21)	118 (19)	101 (17)	83 (98)	85 (100)
Microsatellite	stable	392 (68)	400 (65)	400 (69)	2 (2)	0 (0)
	instable	11 (2)	19 (3)	15 (2)	0 (0)	0 (0)
	no data	176 (30)	198 (32)	169 (29)	83 (98)	85 (100)

Table 2.2. Clinicopathological data of patients by trial arm. Data shown for patients that were genotyped and passed quality control. Percentages shown in parentheses. Response defined as complete or partial response as outlined in RECIST 1.0 guidelines. Non-response defined as stable or progressive disease. OS=overall survival, QC=quality control, Age=age at randomisation, n/k=not known, FOLFOX=oxaliplatin and intravenous 5-FU, folinic acid (leucovorin), XELOX=intravenous oxaliplatin and orally administered capecitabine.

2.3.1.3 Germline DNA analyses

DNA was extracted from blood samples from 2,244 patients by conventional methods and genotyped using Affymetrix Axiom Arrays (Al-Tassan et al. 2015). The genotyping quality was tested using duplicate DNA samples with >99% concordance. Prediction of untyped SNPs was carried out using IMPUTE2 v2.3.0 (Howie et al. 2009) based on data from the 1000 Genomes Project as reference (Howie et al. 2011; Altshuler et al. 2015) (total number of SNPs following imputation = 47,368,871).

2.3.1.4 Germline genotyping quality control

Pre-GWAS QC of the genotyping data was completed in line with current recommendations (Marees et al. 2018b). Individuals were excluded from analysis if they failed one or more of the following thresholds: overall successfully genotyped SNPs <99% (n=122), discordant sex information (n=8), low heterozygosity (inbreeding

coefficient >0.2, n=0), classed as out of bounds by Affymetrix (n=30), duplication or cryptic relatedness (proportion identical by descent >0.1, n=4), and evidence of non-white European ancestry by PCA-based analysis (n=130). After QC, genotype data was available on 1,950 patients (**Figure 1.7**). SNPs that reside in established long range LD regions, such as the major histocompatibility complex region, were removed as they can bias the results of PCA. SNPs were removed if they had INFO score (calculated in SNPTEST) <0.8 (n=29,116,015), missingness >2% (n=3,534,993) or HWE exact test (Wigginton et al. 2005) $P<1.0x10^{-6}$ (n=47), leaving 14,717,816 SNPs for analysis. MAF filtering was considered based upon the available sample size for each particular analysis.

2.3.1.5 Somatic tumour DNA analyses

Two thousand one hundred and eighty-four formalin-fixed, paraffin embedded (FFPE) tumour samples were screened for *KRAS* (codons 12, 13 and 61), *NRAS* (codons 12 and 61), *BRAF* (codons 594 and 600) and *PIK3CA* (codons 542, 545, 546 and 1,047) mutations using Pyrosequencing and Sequenom technologies (Smith et al. 2013). Microsatellite instability (MSI) status in tumours was determined using the markers BAT-25 and BAT-26 (**Table 2.2**).

Overall, *KRAS* mutations (G12A, G12C, G12D, G12V, G12R, G12S, G13C, G13D, G13S, G13R, Q61H, Q61L, Q61R and 5 remained uncharacterised) were identified in 863/2157 (40.0%), *NRAS* mutations (G12C, G12D, G12V, G13D, G13R, Q61H, Q61K, Q61L, Q61H, Q61R and one remained uncharacterised) in 84/2092 (4.0%), *BRAF* mutations (D594G and V600E) in 143/1581 (9.0%) and *PIK3CA* mutations (E542K,

E545K, Q546K, H1047L and H1047R) in 189/1442 (13.1%) CRCs. MSI was detected in 45/1239 (3.6%) CRCs.

2.3.1.6 Survival outcomes

Patients from COIN and COIN-B are combined for survival analyses since there was no evidence of heterogeneity in OS between patients when analysed by trial arm (*P*=0.40; Cochran Q test: *P*=1.0, I^2 test: *P*=0.74), trial (*P*=0.49), cetuximab use (*P*=0.41) or type of chemotherapy received (*P*=0.60; **Figure 2.2**).

2.3.1.7 Response to treatment

Assessment of response was performed at 12 weeks; response was defined as complete or partial response using RECIST 1.0 guidelines and no response was defined as stable or progressive disease.

2.3.2 Study of Colorectal Cancer in Scotland (SOCCS)

The SOCCS trial (1999-current) (Theodoratou et al. 2007; He et al. 2019) aims to recruit 10,000 people from Scotland with CRC by 2026 (ethics approval number MREC/01/0/5 obtained from the MultiCentre Research Ethics committee for Scotland). All patients have a confirmed diagnosis of adenocarcinoma of large bowel epithelium, are genotyped using Illumina HumanHap300, HumanHap240S or Illumina iSelect custom panel arrays and imputed using the 1000 Genomes Project (Howie et al. 2011) as reference (imputation score >0.3 used to select SNPs for analysis) (Tenesa et al. 2008; Theodoratou et al. 2018). Following QC, 5,675 patients (1,358 CRC specific

deaths) of which 784 had stage IV CRC (522 deaths) were made available for this study.

2.3.3 International Survival Analysis in Colorectal cancer Consortium (ISACC)

16,964 patients (4,010 deaths) of which 1,847 had stage IV CRC (1,448 deaths) were made available from ISACC which comprised of 15 studies: the Cancer Prevention Study-II (CPS-II) (Calle et al. 2002), the German Darmkrebs: Chancen der Verhutung durch Screening Study (DACHS) (Brenner et al. 2011; Brenner et al. 2012), the Diet Activity and Lifestyle Study (DALS) (Slattery et al. 1997; Slattery et al. 2003), the Early Detection Research Network (EDRN) (Srivastava and Wagner 2020), the Swedish population of the European Prospective Investigation into Cancer (EPIC) (Riboli and Kaaks 1997), the Health Professionals Follow-up Study (HPFS) (Rimm et al. 1991), the Melbourne Collaborative Cohort Study (MCCS) (Giles and English 2002), the Nurses' Health Study (NHS) (Belanger et al. 1980; Colditz et al. 1997), the N9741 clinical trial (Goldberg et al. 2004), the Physician's Health Study (PHS) (Steering-Committee 1989), the Prostate, Lung, Colorectal, and Ovarian Study (PLCO) (Gohagan et al. 2000; Prorok et al. 2000), the UK Biobank (UKB; Section 2.3.5), the VITamins And Lifestyle Study (VITAL) (White et al. 2004), the Women's Health Initiative (WHI) (Anderson et al. 1998), and four Colon Cancer Family Registry (CCFR) sites: Seattle, Ontario, Australia, and the Mayo Clinic (Newcomb et al. 2007). Study participants included individuals of European genetic ancestry diagnosed with CRC and with available genotyping and CRC-specific survival data. All participants provided informed consent for genetic testing, and all studies were approved by their respective Institutional Review Boards.

2.3.4 The UK Biobank

The UK Biobank (UKB) is a prospective cohort study providing deep genetic and phenotypic data on approximately 500,000 individuals (Bycroft et al. 2018). Participants were all from the United Kingdom and aged between 40 and 69. The phenotypic and medical databases are linked to electronic health records as well as the death and cancer registers. Patients also gave blood, urine, and saliva samples, underwent physical activity monitoring, heart and lung function tests, physical measurements, various imaging procedures and completed extensive questionnaires to collect socio-demographic and lifestyle information. UKB participants were selected for this study if their earliest cancer diagnosis (fields 40005.0.0 to 40005.16.0) was an ICD10 code for tumours in the colon or rectum (fields 40006.0.0 to 40006.16.0). Survival time was calculated as time from diagnosis of CRC to date of death (fields 40000.0.0/40000.1.0). The censoring date for survival time was the 28th of February 2021 (the date the death registry data was collected by UKB and later distributed to researchers in August 2021).

2.3.4.1 Genetic data

Whole-genome germline genotyping was completed for 488,377 participants using two closely related arrays, the UK BiLEVE array (807,411 markers) and the UK Biobank Axiom array (825,927 markers) which had a 95% content overlap. SNPs were imputed to >90 million using the Haplotype Reference Consortium (McCarthy et al. 2016), UK10K + 1000 genomes project (Chou et al. 2016) reference panels. Our work

was carried out under project application number 65833 and used participants from both genotyping arrays.

2.3.4.2 Germline genotyping quality control

Following the UK Biobank's own QC procedures for genotyping quality, 487,409 participants had genotyping data available for download. Pre-GWAS QC of the genotyping data was completed in line with current recommendations (Marees et al. 2018b) using the Hawk HPC. Individuals were excluded from analysis if they failed one or more of the following thresholds: overall successfully genotyped SNPs <99% or low heterozygosity (inbreeding coefficient >0.2, n=377), duplication or cryptic relatedness (KING-kinship coefficient >0.0442 for up to third degree cousins, n=73,321), and evidence of non-white European ancestry by PCA-based analysis (n=78,312). After QC, genotype data was available on 335,399 participants. SNPs were removed if they had INFO score (calculated in SNPTEST) <0.8 or MAF<0.01 (n=83,530,907), missingness >5% (n=637,144) or Hardy-Weinberg equilibrium exact test (Wigginton et al. 2005) $P<1.0x10^{-6}$ (n=73,522), leaving 8,854,050 SNPs for analysis. Further MAF filtering was considered per analysis.

2.3.5 The Genotype-Tissue Expression (GTEx) project

The GTEx project version 8 database (Carithers and Moore 2015; null et al. 2020), was used to identify cis eQTL. The database includes expression data for individual genes from 49 tissues linked to genotype for 838 donors aged 20-79 years old. Of these, 84.6% were white, 12.9% African American, 1.3% Asian, 0.2% American Indian with the remaining donors having unknown heritage. eQTL were annotated by

inputting SNP rs ID's into the 'By variant or rs ID' field on the GTEx portal (https://gtexportal.org/home/). Further information on the GTEx project sequencing and eQTL identification methodologies can be found in their documentation: https://gtexportal.org/home/documentationPage.

2.3.6 The Cancer Genome Atlas (TCGA)

The TCGA dataset (Cancer Genome Atlas Research et al. 2013), available at https://portal.gdc.cancer.gov/, contains molecular characterisation for over 20,000 primary cancer samples across 33 cancer types, including genomic, epigenomic, transcriptomic, and proteomic data. Methylation array data collected using the Illumina human methylation 450 platform was downloaded from the TCGA data repository, containing beta coefficients for methylation levels at each of 485,578 CpG islands across the genome.

2.3.7 The Human Protein Atlas (THPA)

THPA (Uhlen et al. 2015) pathology section contains association information between the survival of approximately 8000 cancer patients (across 17 major cancer types) and genome wide RNA expression levels (Uhlen et al. 2017). Anonymised tissue samples and survival data were collected from the TCGA project from the initial release of Genomic Data Commons (GDC) on June 6, 2016. RNA-seq data for 20,090 genes were reported as a median number of fragments per kilobase of exon per million reads (FPKM) generated by TCGA. Available at https://www.proteinatlas.org/.

2.4 Statistical analyses

2.4.1 Survival analyses

Survival outcomes were assessed by univariate and multivariate Cox proportional-hazards models or log-rank test. Visualisation of survival data included Kaplan-Meier and forest plots produced by the R packages *survminer* and *ggplot2*.

2.4.2 Dimensionality reduction of regression covariates

With a small sample size there is a risk of overfitting in the regression models when including many prognostic clinicopathological factors as covariates. To capture the information observed in each of the prognostic clinicopathological factors whilst reducing the dimensionality of the data, PCA was performed using the *psych* R package. A threshold of 70% total variance of the factors explained by their first principal components was used to select the number of principal components to include as covariates per analysis (Jolliffe and Cadima 2016) (**Figure 2.3**).

2.4.3 Genome wide association study

Linear and binary variables were analysed using linear (*--linear* command) and logistic (*--logistic* command) regressions, respectively, in PLINK version 2.0 (Chang et al. 2015). Censored time-to-event variables, including OS, were analysed using the *plinkCoxSurv* command from the *gwasurvivr* R package (Rizvi et al. 2019).

Univariate models consisted of SNP genotype, recorded as 0,1 or 2 for number of copies of the genotyped or imputed allele and the continuous or binary outcome

variable. Multivariate models also included linear and binary variables as covariates recorded directly or as principal components (Section 2.4.2).

Genome wide significance threshold was $P<5.0x10^{-8}$, and the threshold for suggestive significance was $P<1.0x10^{-5}$. GWAS summary statistics were visualised using the *qqman* R package (Turner 2018).

2.4.4 Power considerations

Statistical power to detect a significant association between survival time variables and SNP genotype was calculated using the *survSNP* R package (Owzar et al. 2012). The effect size, MAF and significance threshold used in the calculation was defined per analysis (**Figure 2.4**).

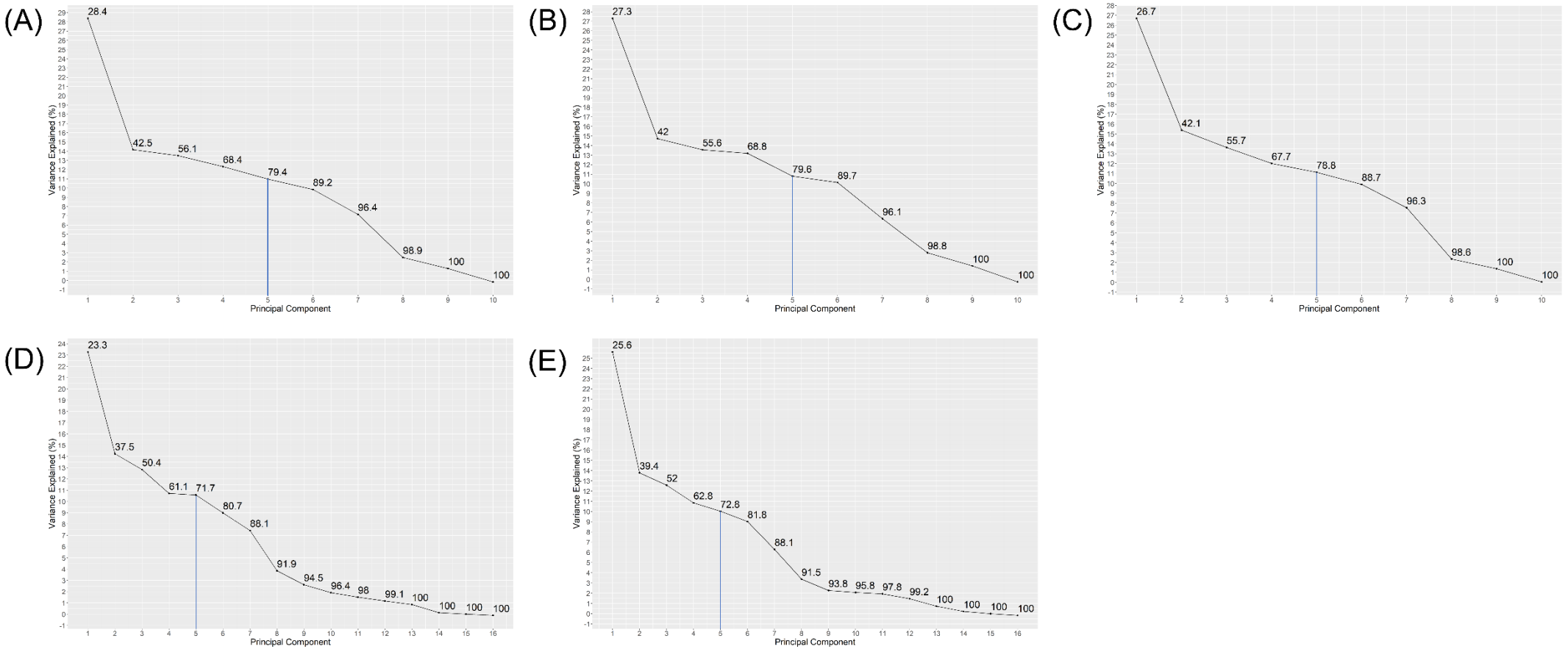


Figure 2.3. Variance of the prognostic clinicopathological factors explained (%) by their first principal components in different cohorts used in this Thesis. (A) 514 patients from COIN and COIN-B with proximal colon tumours, (B) 493 patients with distal colon tumours, (C) 892 patients with rectal tumours, (D) 694 patients with MAPK-activated CRC and (E) 581 patients with wild-type CRC. To capture the information of the clinicopathological factors whilst reducing dimensionality of the regression models a cumulative variance explained (labelled above each point) threshold of 70% was set for inclusion of principal components in the models (annotated in blue). See Chapters 3-6 for details on the clinicopathological factors included in each analysis.

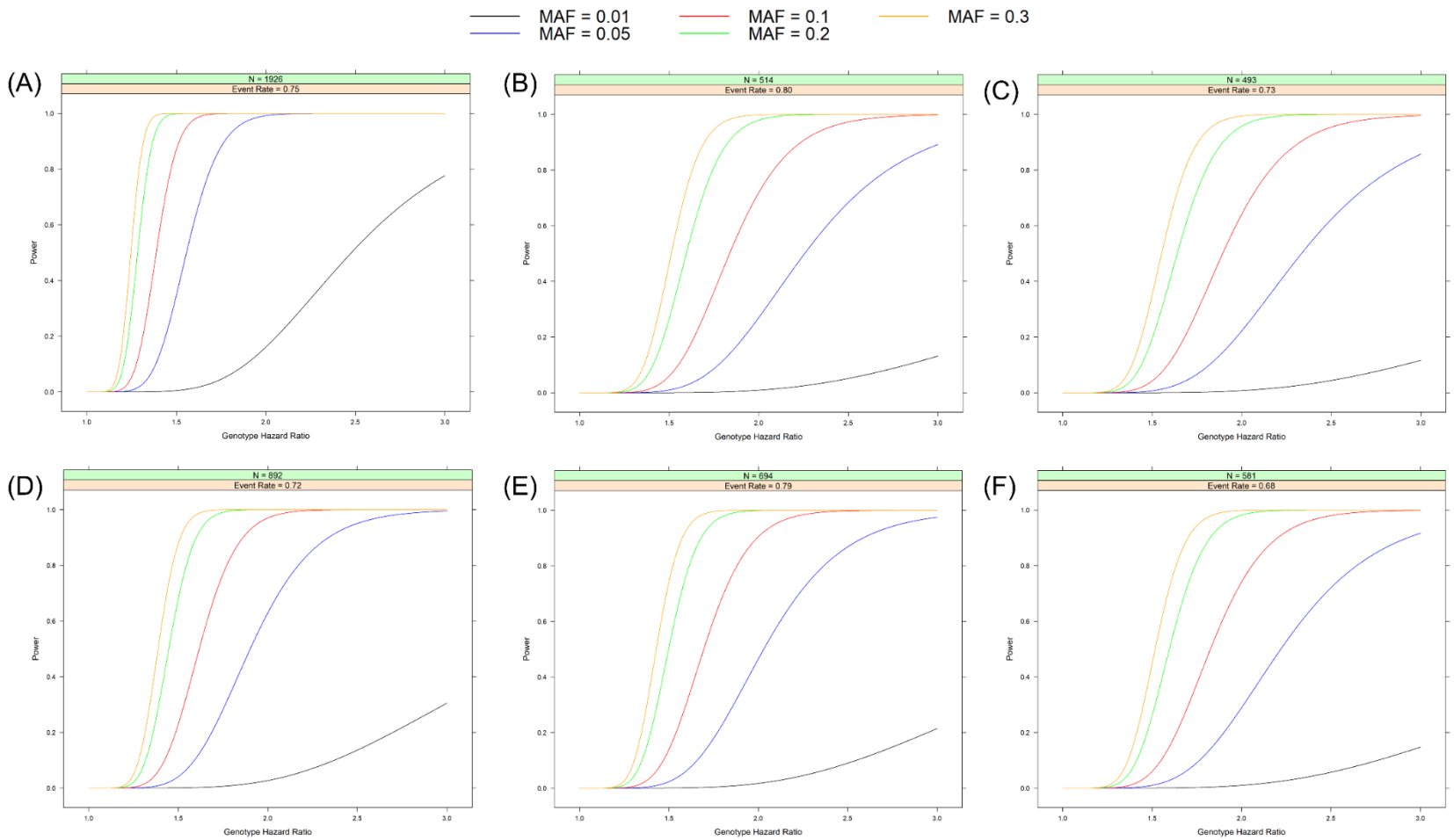


Figure 2.4. Observable hazard ratio per SNP against statistical power for Cox proportional-hazards models in different cohorts used in this thesis: (A) 1926 patients form COIN and COIN-B, (B) 514 patients with proximal colon tumours, (C) 493 patients with distal colon tumours, (D) 892 patients with rectal tumours, (E) 694 patients with MAPK-activated CRC and (F) 581 patients with wild-type CRC. The statistical power can be seen for SNPs at minor allele frequencies ranging from 0.01 to 0.30.

2.4.5 Gene-based and gene-set analyses

Gene and gene-set analyses were performed using MAGMA (de Leeuw et al. 2015) versions 1.07 and 1.09b (https://ctg.cncr.nl/software/magma). SNPs were annotated to genes (including those 35 kilobases before the genes transcription zone and 10 kilobases after) using the *--annotate* command and the gene locations from hg19 build 37.3. SNP *P*-values, taken from the GWAS summary statistics, were assessed with the LD between them using the *multi=snp-wise* and *--gene-model* commands. This model takes advantage of the sum of the -log(*P*) for all SNPs, as well as the top SNP associations within each gene, to assess the association of their constituent genes. A Bonferroni corrected *P*-value threshold of $P<2.5x10^{-6}$ was used to account for 20,000 independent tests (Kiezun et al. 2012).

Genes were annotated to approximately 8000 sets by gene-ontology terms (Ashburner et al. 2000). A competitive model *(--set-result* command) was used to assess each gene-set's association with the outcome variable. The null hypothesis for a competitive test states that each gene in a given gene-set is not more associated with the outcome variable than the other genes in the dataset and is therefore more conservative than a self-contained test. *P*-values were adjusted for false discovery rate (FDR) to produce adjusted *q*-values using the *qvalue* R package (Storey et al. 2021) and significance set at $q<0.05$.

2.4.6 Transcriptome wide association study (TWAS)

Imputation of GReX was completed using the GTEx v8 whole-blood MASHR-based model (downloaded from https://predictdb.org/post/2021/07/21/gtex-v8-models-on-

eqtl-and-sqtl/) and 'Predict.py' script available in the PrediXcan software (Gamazon et al. 2015b). The MASHR-based eQTL models used fine-mapped variants with biological evidence of potential effects on gene expression levels and estimated their effect size in 49 tissues using the GTEx v8 dataset as reference (Barbeira et al. 2021). Individual-level GReX levels were then tested for associations with OS using Cox proportional-hazards models in R.

2.5 Other bioinformatic analyses

2.5.1 LocusZoom plots

LocusZoom (Willer et al. 2010a) was used to produce regional association plots of GWAS summary statistics. The LD of SNPs adjacent to the sentinel SNP (expressed as an r^2 value), recombination rate (in centimorgans per magabase) and genes in the area (relative to hg19) are plotted.

2.6 Study design

All analyses were performed retrospectively with sample size determined by recruitment of patients into the individual study cohorts. No stratification for disease stage was made in either the COIN or COIN-B trial cohort, due to all patients having advanced CRC (stage IV), or the UK Biobank cohort due to missing data on disease stage.

Chapter 3: Genome-wide search for determinants of survival in 1,926 patients with advanced colorectal cancer with follow-up in over 22,000 patients

3.1 Introduction

Clinical stage, which combines depth of tumour invasion, nodal status and distant metastasis (Walther et al. 2009), is currently the only routinely used marker of survival from CRC. Other factors thought to influence patient prognosis include lifestyle (Haydon et al. 2006; Reeves et al. 2007), systemic inflammatory response (Leitch et al. 2007), immunologic microenvironment (Galon et al. 2006) and the patient's germline and the tumour's somatic profile (Popat et al. 2005; Walther et al. 2008). The search for inherited prognostic factors has primarily focussed on candidate genes and SNPs that function in pharmacological pathways (Marcuello et al. 2004; Dotor et al. 2006), influence tumour progression (Kim et al. 2008) or alter disease risk (Dai et al. 2012; Phipps et al. 2012; Abuli et al. 2013; Garcia-Albeniz et al. 2013; Takatsuno et al. 2013; Morris et al. 2015). However, apart from rs9929218 in *CDH1*, most reported SNP associations have not been independently replicated (Smith et al. 2015).

GWAS have been used successfully to identify 205 CRC-susceptibility alleles in the European and east Asian populations, with a further 53 risk loci identified from transcriptomic and methylomic analyses (Fernandez-Rozadilla et al. 2023). To-date, the application of GWAS-based strategies for the identification of alleles influencing survival from CRC has been limited. SNPs near to *ELOVL5* and *DCC* have been

associated with survival in a restricted discovery analysis but not replicated in follow-up (Phipps et al. 2016) and SNPs in *FHIT*, *EPHB1* and *MIR7515* have been associated with time to metastasis but await independent replication (Penney et al. 2019). Here, I report a GWAS of survival in 1,926 patients with advanced CRC from COIN and COIN-B with follow-up of promising SNP-associations in over 22,000 CRC patients from clinical trial and population-based studies.

3.2 Materials and methods

3.2.1 Patients and samples

Of the 2,671 patients recruited to COIN and COIN-B, 1,948 had germline genotyping and survival data available. The minimum SNP MAF was set at 5% leaving 2.9 million SNPs for analysis. See Chapter 2, Section 2.3 for full details on patients, DNA extraction, genotyping and QC.

3.2.2 Statistical analyses

Somatic and clinicopathological factors available in COIN and COIN-B (trial, trial arm, cetuximab status, sex, age, mutation status at *KRAS*, *BRAF*, *NRAS* and *PIK3CA*, MSI status, WHO performance status, resection status of the primary tumour, site of primary tumour, surface area, white blood cell [WBC] count, alkaline phosphatase level, platelet count, chemotherapy regimen, chemotherapy dose, radiotherapy, number of metastatic sites, metastases in the liver, lung, lymph nodes, peritoneum and other sites, time to metastases, synchronous or metachronous metastases, creatinine clearance, glomerular filtration rate and carcinoembryonic antigen [CEA] level) were analysed for their effects on OS using either linear or logistic models. For those shown to be prognostic after Bonferroni correction ($P<1.6x10^{-3}$, n=31 tests), we performed a GWAS for each factor to identify potential SNPs with pleiotropic effects on survival. Lead SNPs at credible independent loci (those with multiple SNPs in the linkage block and that reached the threshold for suggestive significance) were tested for their effects on OS.

We carried out a multivariate GWAS of OS under an additive model for patients in COIN and COIN-B using prognostic covariates that were available in the majority of patients (22 patients excluded, leaving 1,926 for analysis). The covariates included were WHO performance status, resection status of the primary tumour, WBC count, platelet count, alkaline phosphatase levels, number of metastatic sites, metastases in the liver, site of primary tumour (encoded as 7 binary variables), surface area of primary tumour, time from diagnosis to metastases, and metachronous versus synchronous metastases. For any SNPs that reached suggestive significance we conducted a sensitivity analysis replacing OS (considered left-truncated at randomisation since randomisation is conditional upon survival from diagnosis) with time from diagnosis to death or end of trial using Cox regressions. To test for differences in association between the two measures of survival, for each SNP we calculated differences in beta-coefficients and standard errors to produce a chi-squared distribution with 1 degree of freedom; from this *P*-values were determined. See Chapter 2, Section 2.3.2.1 for details on measurement of response to treatment.

Gene and gene-set analysis was completed on the summary statistics from the association analysis to identify genes containing significant numbers of highly associated SNPs and significantly enriched gene-sets (Chapter 2, Section 2.4.5).

3.2.3 Bioinformatic analyses

See Chapter 2, Sections 2.4.3, 2.5.1 and 2.3.5 for details on GWAS analysis, LocusZoom plots and eQTL analyses, respectively. THPA (Chapter 2, Section 2.3.7) was used to find associations between *ERBB4* expression levels in colorectal tumours

and survival in 438 patients with colon adenocarcinomas. Samples were classified as high expression using a threshold of FPKM>0 as per THPA recommendations.

3.2.4 Replication series

Independent replication of lead SNPs at 17 loci showing suggestive evidence of an association with OS in COIN and COIN-B was performed in two independent patient series:

(i) SOCCS (Chapter 2, Section 2.3.2) - 5,675 patients (1,358 CRC specific deaths) of which 784 had stage IV CRC (522 deaths). We considered CRC-specific survival, assigned as time from diagnosis to death from CRC and applied a Cox proportional-hazards model and corrected for age, sex and AJCC stage.

(ii) ISACC (Chapter 2, Section 2.3.3) - 16,964 patients (4,010 deaths) of which 1,847 had stage IV CRC (1,448 deaths). We considered disease-specific survival, applied a Cox-proportional hazards model and corrected for age at diagnosis, sex, genotyping batch, study and the first 5 principal components of genetic ancestry.

3.2.5 Meta-analyses of the follow-up cohorts

Meta-analyses were performed using the inverse variance based method in the METAL software package (Willer et al. 2010b). $P<0.05$ was considered significant for replication of the findings in the discovery cohort.

3.3 Results

3.3.1 Effect of clinicopathological factors on OS

We determined the influence of clinicopathological factors and somatic mutation status on OS in 1,948 patients from COIN and COIN-B. We found that *KRAS* and *BRAF* mutation status, MSI status, platelet count, CEA levels, WHO performance status, resection status of the primary tumour, WBC count, alkaline phosphatase levels, number of metastatic sites, metastases in the liver, lymph nodes and peritoneum, site and surface area of the primary tumour, time from diagnosis to metastases and metachronous versus synchronous metastases were all associated with OS after Bonferroni correction (**Table 3.1**).

3.3.2 GWAS of significant clinicopathological factors

We considered whether SNPs associated with these factors might influence OS and conducted independent GWAS for each factor (n=16). One SNP was associated with WBC count (rs142358223 at 16p13.3, beta coefficient [beta]=1.36, standard error [SE]=0.25, P=3.5x10^{-8}) and two SNPs with CEA levels (rs17418475 at 1p21.2, beta=932.53, SE=163.05, P=1.3x10^{-8} and rs72870425 at 2q24.2, beta=1196.53, SE=211.27, P=1.8x10^{-8}). We tested rs142358223, rs17418475, rs72870425 and 133 lead SNPs from other suggestive loci for their effects on OS, however, none were significant after adjustment for multiple testing (P<3.7x10^{-4}).

Clinicopathological factor	Description	No. genotyped	Overall survival *P*
Trial	COIN or COIN-B	1948	0.49
Arm	Trial arm (A to E)	1948	0.41
Cetuximab	Cetuximab use (yes/no)	1948	0.41
Sex	Sex of patient	1948	2.9×10^{-3}
Age	Age of patient at recruitment (years)	1948	0.76
KRAS	Somatic *KRAS* mutation (yes/no)	1625	**7.1×10^{-6}**
BRAF	Somatic *BRAF* mutation (yes/no)	1581	**1.5×10^{-13}**
NRAS	Somatic *NRAS* mutation (yes/no)	1594	0.49
MSI	Somatic microsatellite instability (yes/no)	1301	**1.9×10^{-5}**
PIK3CA	Somatic *PIK3CA* mutation (yes/no)	1478	0.25
WHO Performance Status	WHO Performance Status rating (0 to 5)	1948	**3.1×10^{-23}**
Resection Status	Primary tumour resected (yes/no/local recurrence)	1948	**1.8×10^{-21}**
Site of Primary Tumour	Primary tumour location	1948	**9.1×10^{-9}**
Surface Area	Surface area of primary tumour	1945	**1.1×10^{-5}**
White Blood Cell Count	White blood cell count ($x10^9$/Litre of blood)	1946	**1.2×10^{-31}**
Alkaline Phosphatase	Alkaline Phosphatase levels (International Units/Litre of blood)	1947	**1.5×10^{-27}**
Platelet Count	Platelet count ($x10^9$/Litre of blood)	1943	**1.7×10^{-29}**
Chemotherapy Regimen	XELOX or FOLFOX based chemotherapy	1948	0.60
Chemotherapy Dose	Intermittent or continuous chemotherapy	1948	0.27
Radiotherapy	Patient received radiotherapy (yes/no)	1948	0.52
Metastatic Sites	Number of separate sites containing metastases	1948	**1.7×10^{-13}**
Liver Metastases	Presence of metastases in the liver (yes/no)	1948	**1.3×10^{-4}**
Lung Metastases	Presence of metastases in the lung (yes/no)	1948	0.53
Nodal Metastases	Presence of metastases in the lymph nodes (yes/no)	1948	**1.5×10^{-3}**
Peritoneal Metastases	Presence of metastases in the peritoneum (yes/no)	1948	**1.6×10^{-7}**
Other Metastases	Presence of metastases elsewhere in the body (yes/no)	1948	**3.4×10^{-5}**
Time to Metastases	Time from primary diagnosis to metastases (days)	1933	**1.7×10^{-7}**
Synchronous or Metachronous	Synchronous or metachronous metastases	1933	**6.0×10^{-8}**
Creatinine Clearance	Volume of blood plasma that is cleared of creatinine per unit time (mL/min)	1744	0.49
Glomerular Filtration Rate	Volume of blood that passes through the glomeruli per unit time (mL/min)	1945	0.32
Carcinoembryonic Antigen Test	Mass of carcinoembryonic antigen per unit of blood (ng/mL)	1518	**2.9×10^{-5}**

Table 3.1. Clinicopathological factors associated with overall survival in COIN and COIN-B (univariate analyses). Significant *P*-values after Bonferroni correction ($P<1.6 \times 10^{-3}$) are highlighted in bold.

3.3.3 Multivariate GWAS of OS

We carried out a multivariate GWAS for OS in 1,926 patients from COIN and COIN-B adjusting for all 11 prognostic covariates (**Figure 3.1**). No detectable genomic inflation was observed (λ=1.08). We had >80% power to detect a HR of 1.3 for SNPs with MAFs ≥0.20 (Chapter 2, Section 2.4.4).

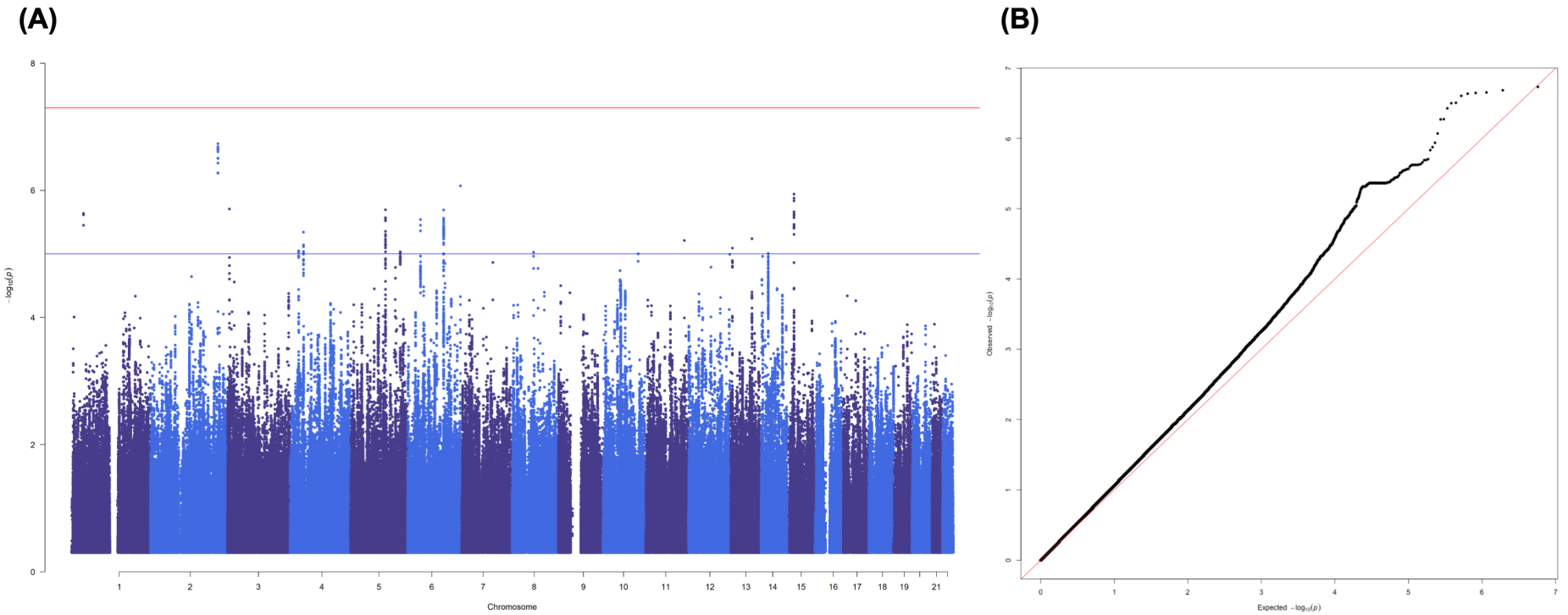


Figure 3.1. Single nucleotide polymorphism (SNP) associations with overall survival (OS) (n=1,926 patients with advanced CRC from COIN and COIN-B). (A) Manhattan plot: SNPs are ordered by chromosome position and plotted against the -$\log_{10}$(*P*) for their association with OS. The red line represents the threshold for genome wide significance (P=5.0x10^{-8}) and the blue line is the threshold for suggestive significance (P=1.0x10^{-5}). Covariates included: World Health Organisation performance status, resection status of the primary tumour, white blood cell count, platelet count, alkaline phosphatase levels, number of metastatic sites, metastases within or outside of the liver, site of primary tumour, surface area of primary tumour, time from diagnosis to metastases and metachronous versus synchronous metastases. **(B)** Quantile-quantile plot: expected -$\log_{10}$(*P*-value), under the null hypothesis of no association between genotype and OS, plotted against observed -$\log_{10}$(*P*-value).

No SNPs reached genome-wide significance. The most significant SNP associated with OS was rs79612564 in *ERBB4* (HR=1.24, 95% CI=1.16-1.32, P=1.9x10^{-7}). Median survival for patients in COIN and COIN-B carrying one minor allele was reduced by 46 days and for those homozygous for the minor allele by 81 days (**Figure 3.2**, **Table 3.2**).

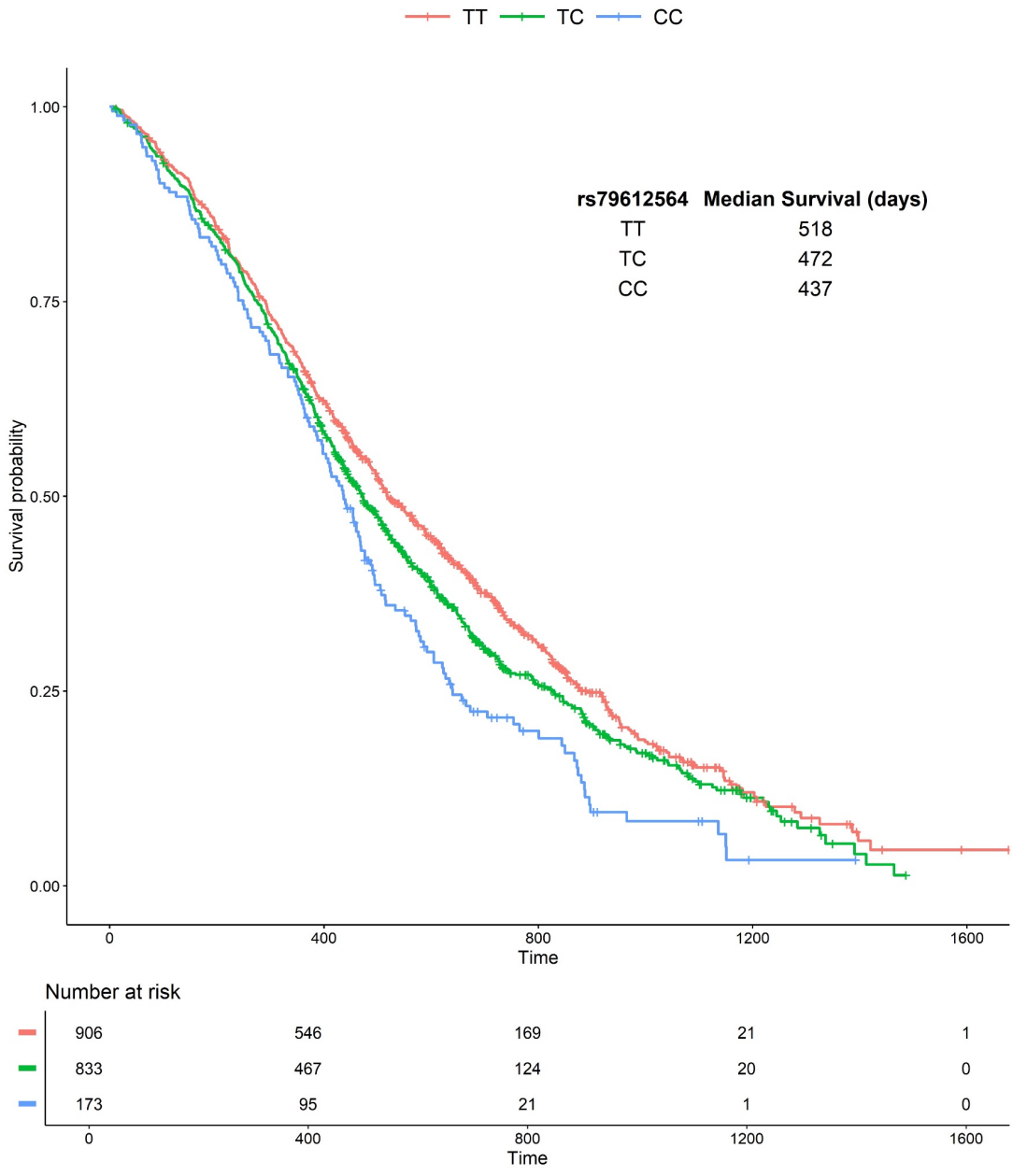


Figure 3.2. Kaplan-Meier plot for rs79612564 genotype in patients with advanced CRC from COIN and COIN-B (n=1,912 patients). Time in days plotted against survival probability for patients homozygous for the major allele (TT), heterozygous (TC) and homozygous for the minor allele (CC). The number of patients still at risk at each time point is shown beneath.

Copies of the minor allele	Median survival	95% CI
0	518	496-572
1	472	441-509
2	437	396-476

Table 3.2. Median survival (days) by rs79612564 genotype for patients in COIN and COIN-B. Copies of the minor allele (C), median survival in days and 95% confidence intervals (CI) for the median are shown.

rs79612564 was not influenced by cetuximab treatment regardless of *KRAS* status (**Figure 3.3**). The prognostic effect appeared to be independent of *KRAS* status and patients carrying at least one rs79612564 minor allele and *KRAS* mutant CRCs had the greatest effect on survival (HR=1.51, CI=1.29-1.77, P=3.7x10^{-7}) (**Figure 3.4**).

In terms of response to oxaliplatin and fluoropyrimidine-based chemotherapy, patients carrying one or more rs79612564 minor alleles showed less response (55.5% for heterozygotes and 55.9% for homozygotes) as compared to patients carrying both major alleles (60.2%), although this did not reach statistical significance (P=0.06) (**Table 3.3**). rs79612564 was not an eQTL.

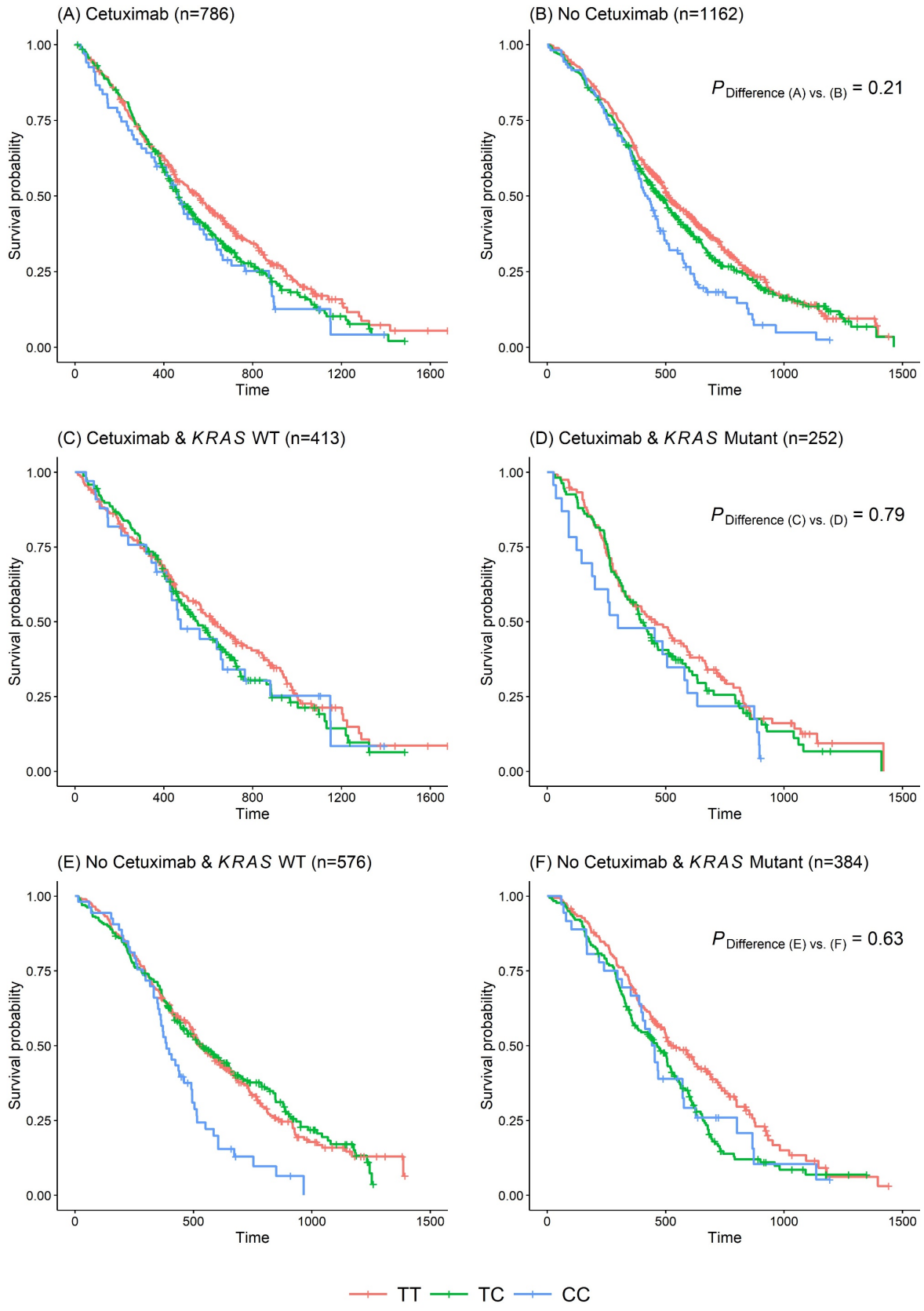


Figure 3.3. Kaplan-Meier plots for rs79612564 genotype in patients treated with and without cetuximab, and by somatic *KRAS* status. Time in days plotted against

survival probability for patients who were homozygous for the major allele (TT), heterozygous (TC) and homozygous for the minor allele (CC) and who (**A**) received cetuximab and (**B**) did not receive cetuximab, irrespective of their *KRAS* status, and who received cetuximab and had *KRAS* WT (**C**) and mutant (**D**) CRCs, and did not receive cetuximab and had *KRAS* WT (**E**) and mutant (**F**) CRCs. *P*-values for the difference in beta coefficients between multivariate Cox-proportional hazards models for rs79612564 against survival time were calculated. WT – wild type.

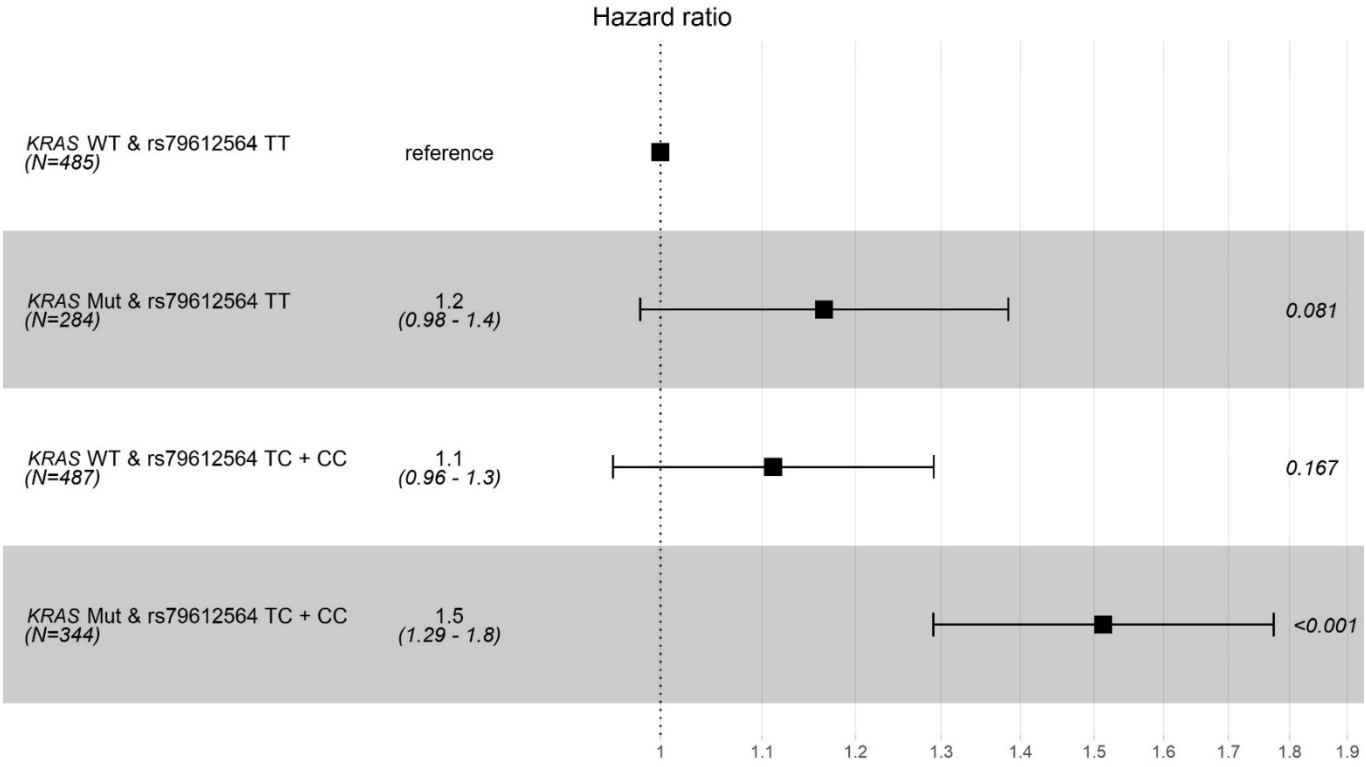


Figure 3.4. Forest plot showing the relationship between *KRAS* mutation status and rs79612564 genotype in patients with advanced CRC from COIN and COIN-B. Hazard ratios, 95% confidence intervals and *P*-values are relative to the reference population who were wild type (WT) for *KRAS* and homozygous for the rs79612564 major allele (TT). Subpopulations had somatically mutated (Mut) *KRAS* +/- rs79612564 minor allele(s).

	rs79612564 genotype		
	TT	**TC**	**CC**
Responders	459	391	85
Non-Responders	303	313	67
% Responders	60.2	55.5	55.9

Table 3.3. Relationship between response to oxaliplatin and fluropyrimidine-based chemotherapy in patients from COIN and COIN-B, and rs79612564 genotype.

rs79612564 had an INFO score of 0.99. We sought independent confirmation of the quality of genotyping and predictive score for this SNP by genotyping rs79612564 directly via KASPar technology. For those samples with both KASPar genotyping and an imputed genotype, we had >99% (1,687/1,703) genotype concordance (**Figure 3.5**).

3.3.4 Other loci of suggestive significance

In total, we identified SNPs at 17 independent loci with suggestive associations with OS (**Table 3.4**, **Figure 3.1**). We conducted a sensitivity analysis for lead SNPs at all 17 loci replacing OS with an alternative measure of survival - time from diagnosis to death or end of trial. There were no significant differences between the two measures of survival for any of the 17 SNPs (*P*=0.46-0.95). rs6568761 at 6q21 (in a gene desert) passed the threshold for genome wide significance with diagnosis to death (HR=0.88, 95% CI=0.78-0.98, $P=4.5x10^{-8}$).

We did not find any significantly associated genes (**Table 3.5**), or gene-sets under competitive analyses (**Table 3.6**) for OS after correction for multiple testing.

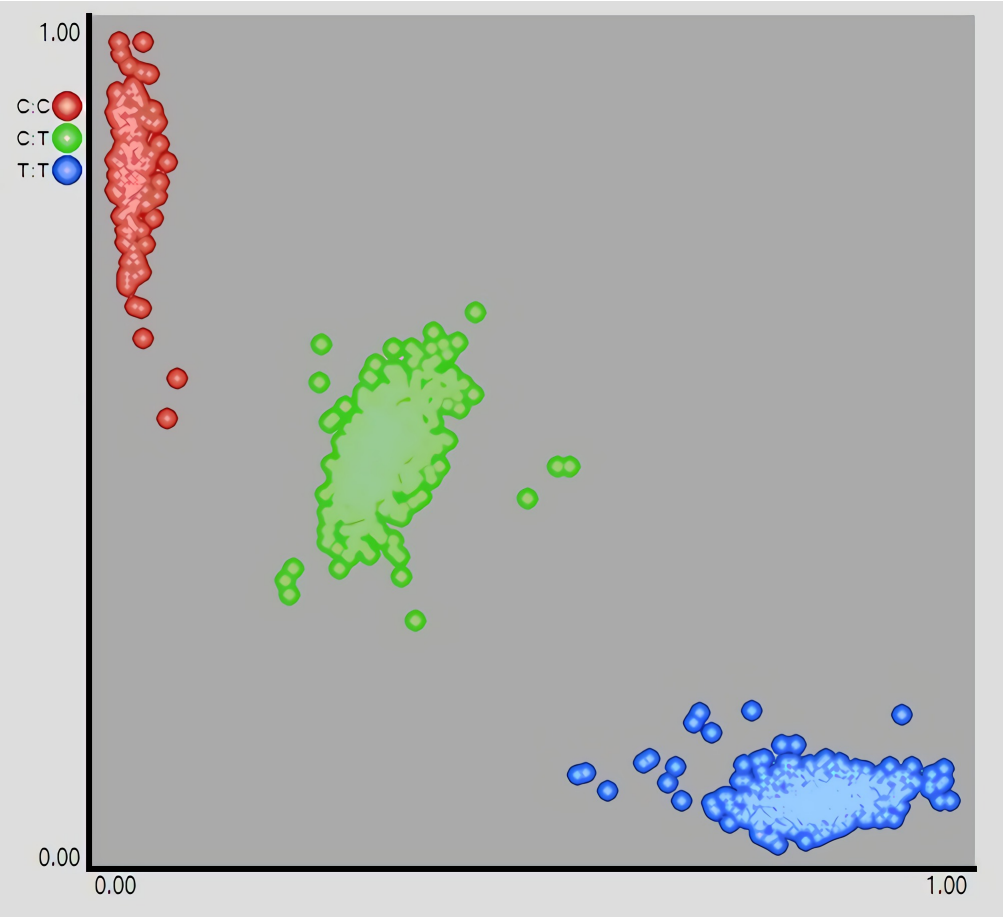


Figure 3.5. Independent assessment of rs79612564 genotyping using KASPar. For those samples with both KASPar genotyping and an imputed genotype, we had >99% (1,687/1,703) genotype concordance: 98.7% (156/158) of samples imputed as homozygous for the minor allele matched that genotype (red), 98.7% (733/743) of samples imputed as heterozygous matched that genotype (green), and 99.5% (798/802) of samples imputed as homozygous for the major allele matched that genotype (blue).

SNP	Locus	Minor Allele	Genes	Overall survival			Diagnosis to death		
				HR	95% CI	*P*	HR	95% CI	*P*
rs79612564	2q34	C	*ERBB4*	1.24	1.16-1.32	$1.9x10^{-7}$	1.08	1.00-1.16	$4.7x10^{-5}$
rs9356458	6q27	A		0.82	0.75-0.90	$9.1x10^{-7}$	0.92	0.85-1.00	$1.1x10^{-5}$
rs9744647	15q14	T	*C145orf51*	1.29	1.18-1.39	$2.0x10^{-6}$	1.11	1.03-1.20	$4.3x10^{-6}$
rs6568761	6q21	G		0.78	0.67-0.88	$2.0x10^{-6}$	0.88	0.78-0.98	**$4.5x10^{-8}$**
rs244509	5q22.1	C	*CAMK4*	0.81	0.73-0.90	$2.0x10^{-6}$	0.91	0.83-0.99	$1.0x10^{-6}$
rs1400673	3p25.1	G		1.35	1.23-1.48	$2.1x10^{-6}$	1.13	1.01-1.25	$1.4x10^{-5}$
rs4653255	1p34.3	A		0.84	0.76-0.91	$2.6x10^{-6}$	0.94	0.86-1.02	$1.1x10^{-4}$
rs2473571	6p21.1	G	*LRFN2*	1.19	1.12-1.27	$3.1x10^{-6}$	1.06	0.98-1.14	$3.5x10^{-4}$
rs9594035	13q31.1	T		0.82	0.73-0.90	$5.4x10^{-6}$	0.92	0.84-1.00	$5.4x10^{-6}$
rs3103204	4p13	T	*ATP8A1, SHISA3*	0.76	0.64-0.88	$5.4x10^{-6}$	0.89	0.78-1.01	$2.5x10^{-5}$
rs11605969	11q24.1	T	*SORL1*	1.26	1.16-1.36	$6.3x10^{-6}$	1.08	0.98-1.18	$3.3x10^{-4}$
rs4411363	13q12.12	G	*TNFRSF19*	1.19	1.12-1.27	$7.8x10^{-6}$	1.06	0.98-1.14	$1.1x10^{-3}$
rs1352374	4p15.2	C		0.82	0.73-0.91	$8.4x10^{-6}$	0.92	0.80-1.03	$1.8x10^{-5}$
rs6983214	8q13.1	T	*C8orf44, C8orf44-SGK3, VCPIP1*	0.83	0.75-0.91	$8.8x10^{-6}$	0.92	0.84-1.00	$4.9x10^{-6}$
rs11744800	5q33.3	C	*ADAM19*	0.82	0.74-0.91	$8.8x10^{-6}$	0.93	0.85-1.01	$3.5x10^{-4}$
rs2050337	10q25.1	G		1.19	1.11-1.26	$9.0x10^{-6}$	1.07	0.99-1.15	$6.5x10^{-5}$
rs7145600	14q21.1	T		0.79	0.69-0.90	$9.5x10^{-6}$	0.91	0.81-1.01	$5.2x10^{-5}$

Table 3.4. Lead single nucleotide polymorphisms (SNPs) from independent loci that reached suggestive significance in multivariate analysis of overall survival (OS) in COIN and COIN-B. Cytogenic band, minor allele, *P*-value, hazard ratio and 95% confidence intervals are shown for OS (time from trial recruitment to death or end of study) and time from diagnosis to death or end of trial. Only rs6568761 reached the threshold for genome-wide significance ($P<5.0x10^{-8}$, highlighted in bold). Genes overlapping with the SNPs attributed to each locus are listed.

Gene Name	Chromosome	Start	Stop	*P*
VCPIP1	8	67532488	67614452	$8.7x10^{-6}$
C8orf44	8	67544787	67607797	$1.2x10^{-5}$
SHISA3	4	42364856	42414504	$1.5x10^{-5}$
MYBL1	8	67464410	67560484	$1.5x10^{-5}$
C8orf44-SGK3	8	67544787	67784257	$1.9x10^{-5}$
LRFN2	6	40349373	40590126	$2.1x10^{-5}$
SGK3	8	67589653	67784257	$2.9x10^{-5}$
SORL1	11	121287912	121514471	$3.1x10^{-5}$
C15orf41	15	36836812	37112449	$7.2x10^{-5}$

Table 3.5. Results for MAGMA gene analysis. All genes with $P<1.0x10^{-4}$ as well as their chromosome, start and stop positions are shown. None reached statistical significance ($P<2.5x10^{-6}$).

GO Term	Gene-Set Name	*P*	*q*
GO:0008219	cell death	$3.0x10^{-5}$	0.076
GO:0012501	programmed cell death	$4.3x10^{-5}$	0.076
GO:0046133	pyrimidine ribonucleoside catabolic process	$3.9x10^{-5}$	0.076
GO:0035774	positive regulation of insulin secretion involved in cellular response to glucose stimulus	$4.7x10^{-5}$	0.076
GO:0071071	regulation of phospholipid biosynthetic process	$2.3x10^{-5}$	0.076
GO:0060390	regulation of SMAD protein signal transduction	$6.8x10^{-5}$	0.092

Table 3.6. Results for MAGMA gene-set enrichment analysis. Gene-ontology (GO) term, full descriptive name, *P*-value, and corrected *P*-value (*q*) are shown. Only sets with $q<0.10$ are presented; none reached statistical significance ($q<0.05$).

3.3.5 Replication analyses

We analysed lead SNPs at all 17 loci in 5,675 patients with CRC from SOCCS and 16,964 patients with CRC from ISACC (**Table 3.7**, **Figure 3.6**). Together, we had >98% power to replicate all 17 SNPs (alpha=0.05). After meta-analysis, no lead SNPs were independently replicated and only rs1352374 and rs2050337 reached nominal significance in SOCCS (**Table 3.7**).

We considered whether the lack of replication of the COIN and COIN-B data might be confounded by patients with differing stages of disease in the follow-up cohorts. We therefore tested the 17 lead SNPs in a subset of 784 patients from SOCCS and 1,847 patients from ISACC with stage IV CRC (**Table 3.8**, **Figure 3.7**). We had >80% power to replicate 16 of the SNPs (for rs3103204 we had 62% power, alpha=0.05). rs79612564 was significant in stage IV patients from SOCCS (P=2.1x10^{-2}) but not in stage IV patients from ISACC (P=0.89, **Table 3.8**). When SOCCS was combined with COIN and COIN-B, rs79612564 reached genome wide significance (HR=1.22, 95% CI=1.15-1.29, P=1.7x10^{-8}), but not when ISACC was also included (HR=1.12, 95% CI=1.06-1.17, P=3.4x10^{-5}).

rs6983214 was significant in the meta-analysis of stage IV patients from SOCCS and ISACC (P=1.2x10^{-3}), however, the direction of effect was opposite to that found in COIN and COIN-B (**Table 3.8**). rs1352374 reached nominal significance in SOCCS (P=3.3x10^{-2}), but not in ISACC. rs2050337 reached nominal significance in the meta-analysis (P=1.1x10^{-2}, **Table 3.8**) with the same direction of effect in all cohorts tested (meta-analysis with COIN and COIN-B included HR=1.13, 95% CI=1.08-1.18, P=1.6x10^{-6}).

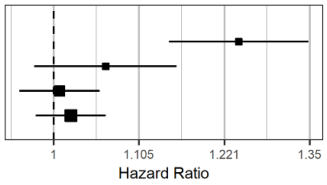
SNP	COIN and COIN-B 1,926 patients (1,435 deaths)		Independent replication						Meta
			SOCCS 5,675 patients (1,358 deaths)			ISACC 16,964 patients (4,010 deaths)			
	HR	95% CI	HR	95% CI	*P*	HR	95% CI	*P*	*P*
rs79612564	1.24	1.16-1.32	1.06	0.98-1.15	0.15	1.01	0.96-1.05	0.77	0.34
rs9356458	0.82	0.75-0.90	1.03	0.95-1.11	0.44	1.00	0.95-1.04	0.87	0.82
rs9744647	1.29	1.18-1.39	1.02	0.90-1.14	0.70	1.02	0.96-1.09	0.45	0.73
rs6568761	0.78	0.67-0.88	0.99	0.88-1.09	0.60	1.01	0.95-1.06	0.86	0.97
rs244509	0.81	0.73-0.90	1.08	0.99-1.16	0.10	1.01	0.96-1.06	0.81	0.30
rs1400673	1.35	1.23-1.48	0.98	0.84-1.12	0.78	1.00	0.92-1.07	0.97	0.87
rs4653255	0.84	0.76-0.91	0.97	0.89-1.04	0.37	0.99	0.95-1.04	0.75	0.47
rs2473571	1.19	1.12-1.27	1.01	0.93-1.09	0.76	0.99	0.95-1.04	0.76	0.91
rs9594035	0.82	0.73-0.90	0.99	0.90-1.08	0.87	1.01	0.96-1.06	0.61	0.72
rs3103204	0.76	0.64-0.88	0.98	0.86-1.10	0.75	0.99	0.93-1.06	0.79	0.70
rs11605969	1.26	1.16-1.36	0.98	0.88-1.09	0.71	1.02	0.95-1.08	0.63	0.82
rs4411363	1.19	1.12-1.27	0.99	0.91-1.07	0.84	1.01	0.96-1.05	0.72	0.84
rs1352374	0.82	0.73-0.91	0.89	0.80-0.98	**$1.5x10^{-2}$**	1.01	0.96-1.06	0.62	0.58
rs6983214	0.83	0.75-0.91	1.07	0.98-1.15	0.13	1.00	0.95-1.05	0.91	0.39
rs11744800	0.82	0.74-0.91	1.04	0.96-1.13	0.33	0.98	0.93-1.03	0.36	0.75
rs2050337	1.19	1.11-1.26	1.09	1.02-1.17	**$2.4x10^{-2}$**	1.01	0.97-1.06	0.60	0.11
rs7145600	0.79	0.69-0.90	1.01	0.91-1.11	0.81	1.01	0.95-1.07	0.79	0.72

Table 3.7. Independent replication of lead SNPs in SOCCS and ISACC. Hazard Ratio, 95% confidence intervals and *P*-value are listed for overall survival (time from trial recruitment to death or end of study) in COIN and COIN-B, and CRC-specific survival (time from diagnosis to death due to CRC) in SOCCS and ISACC. Nominally significant *P*-values are highlighted in bold.

All Stages

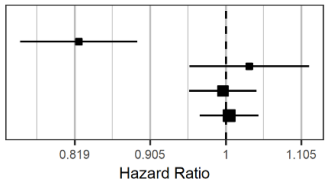
rs79612564

Replication	Study	P	HR	95% CI
	COIN/COIN-B	$1.9x10^{-7}$	1.24	(1.16-1.32)
	SOCCS	0.15	1.06	(0.98-1.15)
	ISACC	0.77	1.01	(0.96-1.05)
	Meta-analysis	0.34	1.02	(0.98-1.06)



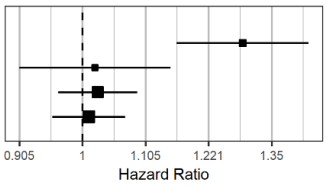
rs9356458

Replication	Study	P	HR	95% CI
	COIN/COIN-B	$9.1x10^{-7}$	0.82	(0.75-0.90)
	SOCCS	0.44	1.03	(0.95-1.11)
	ISACC	0.87	1.00	(0.95-1.04)
	Meta-analysis	0.82	1.00	(0.97-1.04)



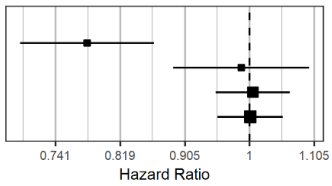
rs9744647

Replication	Study	P	HR	95% CI
	COIN/COIN-B	$2.0x10^{-6}$	1.29	(1.18-1.39)
	SOCCS	0.70	1.02	(0.90-1.14)
	ISACC	0.45	1.02	(0.96-1.09)
	Meta-analysis	0.73	1.01	(0.95-1.07)



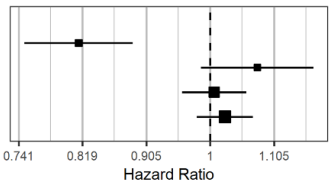
rs6568761

Replication	Study	P	HR	95% CI
	COIN/COIN-B	$2.0x10^{-6}$	0.78	(0.67-0.88)
	SOCCS	0.60	0.99	(0.88-1.09)
	ISACC	0.86	1.01	(0.95-1.06)
	Meta-analysis	0.97	1.00	(0.95-1.05)



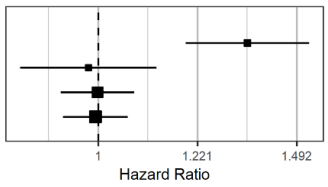
rs244509

Replication	Study	P	HR	95% CI
	COIN/COIN-B	$2.0x10^{-6}$	0.81	(0.73-0.90)
	SOCCS	0.10	1.08	(0.99-1.16)
	ISACC	0.81	1.01	(0.96-1.06)
	Meta-analysis	0.30	1.02	(0.98-1.07)



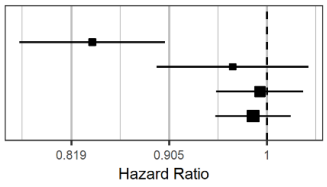
rs1400673

Replication	Study	P	HR	95% CI
	COIN/COIN-B	$2.1x10^{-6}$	1.35	(1.23-1.48)
	SOCCS	0.78	0.98	(0.84-1.12)
	ISACC	0.97	1.00	(0.92-1.07)
	Meta-analysis	0.87	0.99	(0.93-1.06)



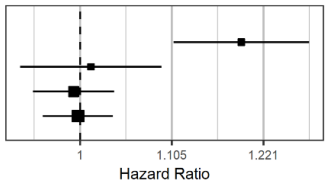
rs4653255

Replication	Study	P	HR	95% CI
	COIN/COIN-B	$2.6x10^{-6}$	0.84	(0.76-0.91)
	SOCCS	0.37	0.97	(0.89-1.04)
	ISACC	0.75	0.99	(0.95-1.04)
	Meta-analysis	0.47	0.99	(0.95-1.02)



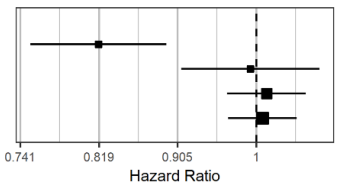
rs2473571

Replication	Study	P	HR	95% CI
	COIN/COIN-B	$3.1x10^{-6}$	1.19	(1.12-1.27)
	SOCCS	0.76	1.01	(0.93-1.09)
	ISACC	0.76	0.99	(0.95-1.04)
	Meta-analysis	0.91	1.00	(0.96-1.04)



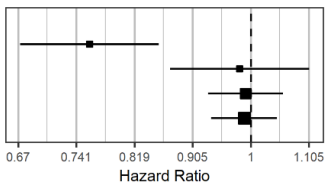
rs9594035

Replication	Study	P	HR	95% CI
	COIN/COIN-B	$5.4x10^{-6}$	0.82	(0.73-0.90)
	SOCCS	0.87	0.99	(0.90-1.08)
	ISACC	0.61	1.01	(0.96-1.06)
	Meta-analysis	0.72	1.01	(0.96-1.05)



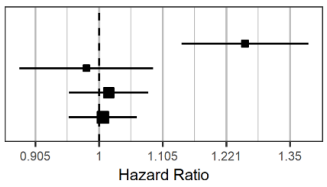
rs3103204

Replication	Study	P	HR	95% CI
	COIN/COIN-B	$5.4x10^{-6}$	0.76	(0.64-0.88)
	SOCCS	0.75	0.98	(0.86-1.10)
	ISACC	0.79	0.99	(0.93-1.06)
	Meta-analysis	0.70	0.99	(0.93-1.05)



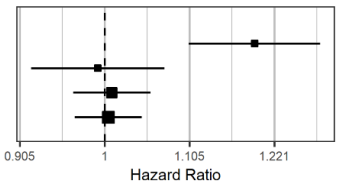
rs11605969

Replication	Study	P	HR	95% CI
	COIN/COIN-B	$6.3x10^{-6}$	1.26	(1.16-1.36)
	SOCCS	0.71	0.98	(0.88-1.09)
	ISACC	0.63	1.02	(0.95-1.08)
	Meta-analysis	0.82	1.01	(0.95-1.06)



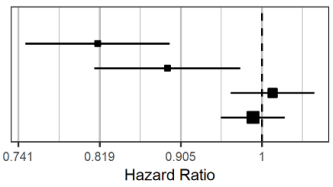
rs4411363

Replication	Study	P	HR	95% CI
	COIN/COIN-B	$7.8x10^{-6}$	1.19	(1.12-1.27)
	SOCCS	0.84	0.99	(0.91-1.07)
	ISACC	0.72	1.01	(0.96-1.05)
	Meta-analysis	0.84	1.00	(0.96-1.04)



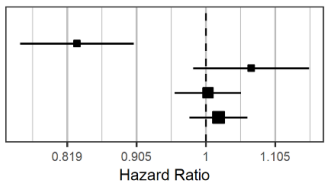
rs1352374

Replication	Study	P	HR	95% CI
	COIN/COIN-B	$8.4x10^{-6}$	0.82	(0.73-0.91)
	SOCCS	$1.5x10^{-2}$	0.89	(0.80-0.98)
	ISACC	0.62	1.01	(0.96-1.06)
	Meta-analysis	0.58	0.99	(0.95-1.03)



rs6983214

Replication	Study	P	HR	95% CI
	COIN/COIN-B	$8.8x10^{-6}$	0.83	(0.75-0.91)
	SOCCS	0.13	1.07	(0.98-1.15)
	ISACC	0.91	1.00	(0.95-1.05)
	Meta-analysis	0.39	1.02	(0.98-1.06)



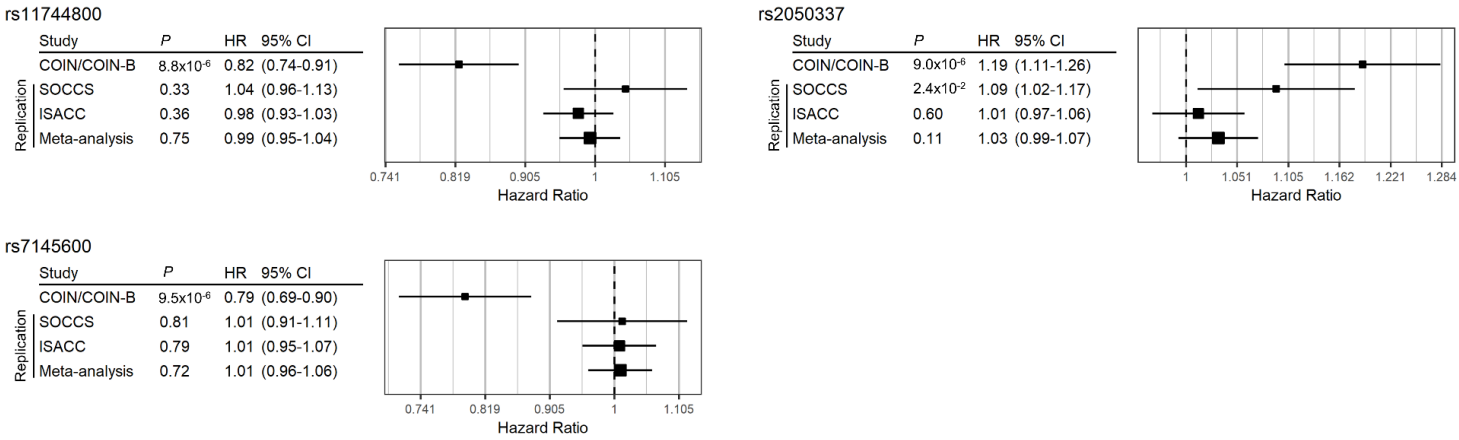


Figure 3.6. Forest plots for lead SNPs at 17 loci identified in COIN and COIN-B and the independent replication cohorts (all stages). *P*-value, Hazard ratio and 95% confidence intervals are listed.

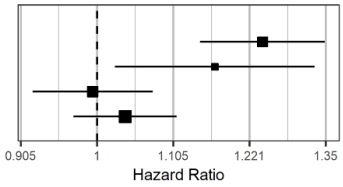
SNP	COIN and COIN-B 1,926 patients (1,435 deaths)		Independent replication						Meta
			SOCCS Stage IV 784 patients (522 deaths)			ISACC Stage IV 1,847 patients (1,448 deaths)			
	HR	95% CI	HR	95% CI	*P*	HR	95% CI	*P*	*P*
rs79612564	1.24	1.16-1.32	1.17	1.04-1.30	**$2.1x10^{-2}$**	0.99	0.92-1.07	0.89	0.28
rs9356458	0.82	0.75-0.90	1.09	0.96-1.21	0.19	-	-	-	-
rs9744647	1.29	1.18-1.39	1.01	0.81-1.21	0.93	0.97	0.86-1.07	0.52	0.82
rs6568761	0.78	0.67-0.88	1.02	0.86-1.17	0.62	1.03	0.93-1.12	0.58	0.56
rs244509	0.81	0.73-0.90	1.08	0.94-1.21	0.30	1.00	0.92-1.09	0.96	0.56
rs1400673	1.35	1.23-1.48	1.03	0.82-1.24	0.78	1.08	0.96-1.21	0.22	0.23
rs4653255	0.84	0.76-0.91	1.00	0.88-1.12	0.97	1.04	0.96-1.11	0.35	0.41
rs2473571	1.19	1.12-1.27	0.99	0.87-1.11	0.86	0.97	0.90-1.05	0.49	0.50
rs9594035	0.82	0.73-0.90	0.96	0.82-1.10	0.57	0.96	0.88-1.05	0.36	0.28
rs3103204	0.76	0.64-0.88	0.89	0.71-1.07	0.19	0.93	0.82-1.03	0.17	0.06
rs11605969	1.26	1.16-1.36	1.12	0.95-1.29	0.18	1.05	0.95-1.15	0.35	0.14
rs4411363	1.19	1.12-1.27	1.03	0.90-1.16	0.65	1.02	0.94-1.10	0.65	0.53
rs1352374	0.82	0.73-0.91	0.85	0.71-0.99	**$3.3x10^{-2}$**	1.00	0.91-1.08	0.99	0.59
rs6983214	0.83	0.75-0.91	1.15	1.02-1.28	**$3.6x10^{-2}$**	1.11	1.03-1.19	**$1.2x10^{-2}$**	**$1.2x10^{-3*}$**
rs11744800	0.82	0.74-0.91	1.03	0.89-1.17	0.72	1.03	0.95-1.12	0.47	0.42
rs2050337	1.19	1.11-1.26	1.08	0.96-1.20	0.22	1.09	1.01-1.17	**$2.7x10^{-2}$**	**$1.1x10^{-2}$**
rs7145600	0.79	0.69-0.90	1.07	0.91-1.23	0.39	0.92	0.82-1.02	0.09	0.32

Table 3.8. Independent replication of lead single nucleotide polymorphisms in patients from SOCCS and ISACC with Stage IV colorectal cancer (CRC). Hazard Ratio, 95% confidence intervals and *P*-value are listed for overall survival (time from trial recruitment to death or end of study) in COIN and COIN-B, and CRC-specific survival (time from diagnosis to death due to CRC) in SOCCS and ISACC. Nominally significant *P*-values are highlighted in bold. *Opposite direction of effect to COIN and COIN-B so not validated. Data for rs9356458, nor any proxies were available for stage IV patients from ISACC.

Stage IV

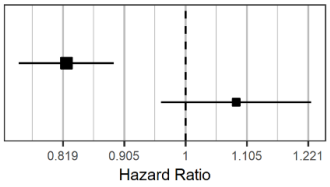
rs79612564

	Study	*P*	HR	95% CI
Replication	COIN/COIN-B	1.9×10^{-7}	1.24	(1.16-1.32)
	SOCCS	2.1×10^{-2}	1.17	(1.04-1.30)
	ISACC	0.89	0.99	(0.92-1.07)
	Meta-analysis	0.28	1.04	(0.97-1.11)



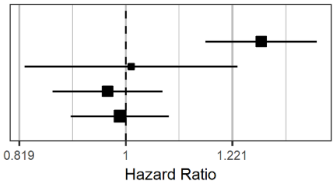
rs9356458

	Study	*P*	HR	95% CI
Replication	COIN/COIN-B	9.1×10^{-7}	0.82	(0.75-0.90)
	SOCCS	0.19	1.09	(0.96-1.21)



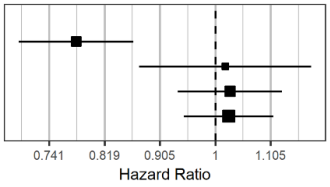
rs9744647

	Study	*P*	HR	95% CI
Replication	COIN/COIN-B	2.0×10^{-6}	1.29	(1.18-1.39)
	SOCCS	0.93	1.01	(0.81-1.21)
	ISACC	0.52	0.97	(0.86-1.07)
	Meta-analysis	0.82	0.99	(0.90-1.08)



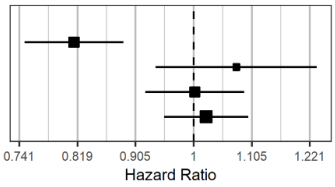
rs6568761

	Study	*P*	HR	95% CI
Replication	COIN/COIN-B	2.0×10^{-6}	0.78	(0.67-0.88)
	SOCCS	0.62	1.02	(0.86-1.17)
	ISACC	0.58	1.03	(0.93-1.12)
	Meta-analysis	0.56	1.02	(0.94-1.10)



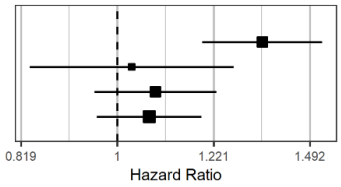
rs244509

	Study	*P*	HR	95% CI
Replication	COIN/COIN-B	2.0×10^{-6}	0.81	(0.73-0.90)
	SOCCS	0.30	1.08	(0.94-1.21)
	ISACC	0.96	1.00	(0.92-1.09)
	Meta-analysis	0.56	1.02	(0.95-1.09)



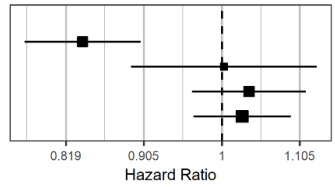
rs1400673

	Study	*P*	HR	95% CI
Replication	COIN/COIN-B	2.1×10^{-6}	1.35	(1.23-1.48)
	SOCCS	0.78	1.03	(0.82-1.24)
	ISACC	0.22	1.08	(0.96-1.21)
	Meta-analysis	0.23	1.07	(0.96-1.18)



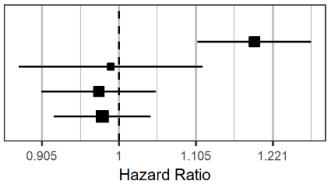
rs4653255

	Study	*P*	HR	95% CI
Replication	COIN/COIN-B	2.6×10^{-6}	0.84	(0.76-0.91)
	SOCCS	0.97	1.00	(0.88-1.12)
	ISACC	0.35	1.04	(0.96-1.11)
	Meta-analysis	0.41	1.03	(0.96-1.09)



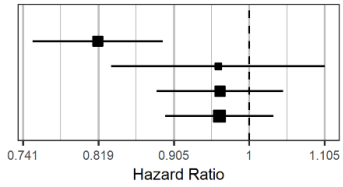
rs2473571

	Study	*P*	HR	95% CI
Replication	COIN/COIN-B	3.1×10^{-6}	1.19	(1.12-1.27)
	SOCCS	0.86	0.99	(0.87-1.11)
	ISACC	0.49	0.97	(0.90-1.05)
	Meta-analysis	0.50	0.98	(0.92-1.04)



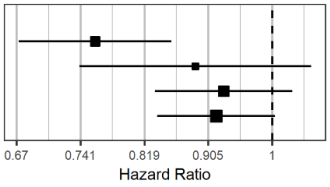
rs9594035

	Study	*P*	HR	95% CI
Replication	COIN/COIN-B	5.4×10^{-6}	0.82	(0.73-0.90)
	SOCCS	0.57	0.96	(0.82-1.10)
	ISACC	0.36	0.96	(0.88-1.05)
	Meta-analysis	0.28	0.96	(0.89-1.03)



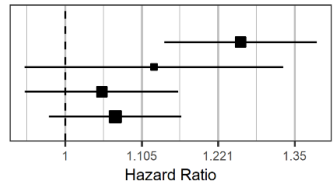
rs3103204

	Study	*P*	HR	95% CI
Replication	COIN/COIN-B	5.4×10^{-6}	0.76	(0.64-0.88)
	SOCCS	0.19	0.89	(0.71-1.07)
	ISACC	0.17	0.93	(0.82-1.03)
	Meta-analysis	0.06	0.92	(0.82-1.01)



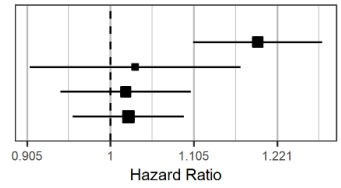
rs11605969

	Study	*P*	HR	95% CI
Replication	COIN/COIN-B	6.3×10^{-6}	1.26	(1.16-1.36)
	SOCCS	0.18	1.12	(0.95-1.29)
	ISACC	0.35	1.05	(0.95-1.15)
	Meta-analysis	0.14	1.07	(0.98-1.15)



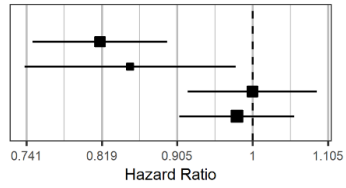
rs4411363

	Study	*P*	HR	95% CI
Replication	COIN/COIN-B	7.8×10^{-6}	1.19	(1.12-1.27)
	SOCCS	0.65	1.03	(0.90-1.16)
	ISACC	0.65	1.02	(0.94-1.10)
	Meta-analysis	0.53	1.02	(0.96-1.09)



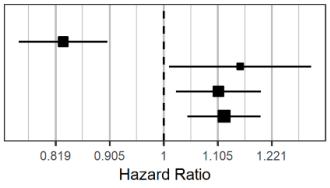
rs1352374

	Study	*P*	HR	95% CI
Replication	COIN/COIN-B	8.4×10^{-6}	0.82	(0.73-0.91)
	SOCCS	3.3×10^{-2}	0.85	(0.71-0.99)
	ISACC	0.99	1.00	(0.91-1.08)
	Meta-analysis	0.59	0.98	(0.90-1.06)



rs6983214

	Study	*P*	HR	95% CI
Replication	COIN/COIN-B	8.8×10^{-6}	0.83	(0.75-0.91)
	SOCCS	3.6×10^{-2}	1.15	(1.02-1.28)
	ISACC	1.2×10^{-2}	1.11	(1.03-1.19)
	Meta-analysis	1.2×10^{-3}	1.12	(1.05-1.19)



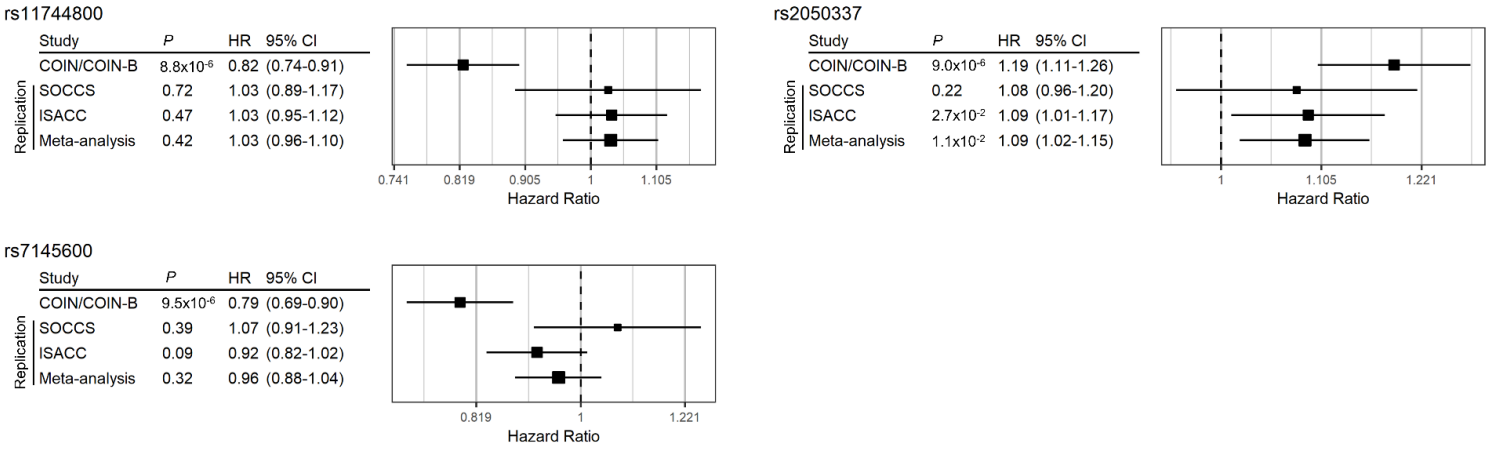


Figure 3.7. Forest plots for lead single nucleotide polymorphisms at 17 loci identified in COIN and COIN-B and the independent replication cohorts (stage IV disease). *P*-value, Hazard ratio and 95% confidence intervals are listed.

3.3.6 Relationship between *ERBB4* expression and survival

We sought additional mechanistic data for a role for *ERBB4* on survival by studying 438 patients with colon adenocarcinomas from THPA. Patients with high *ERBB4* expression in their tumours had worse survival (Cox-regression HR=1.50, 95% CI=1.10-1.90, P=4.6x10^{-2}, **Figure 3.8**).

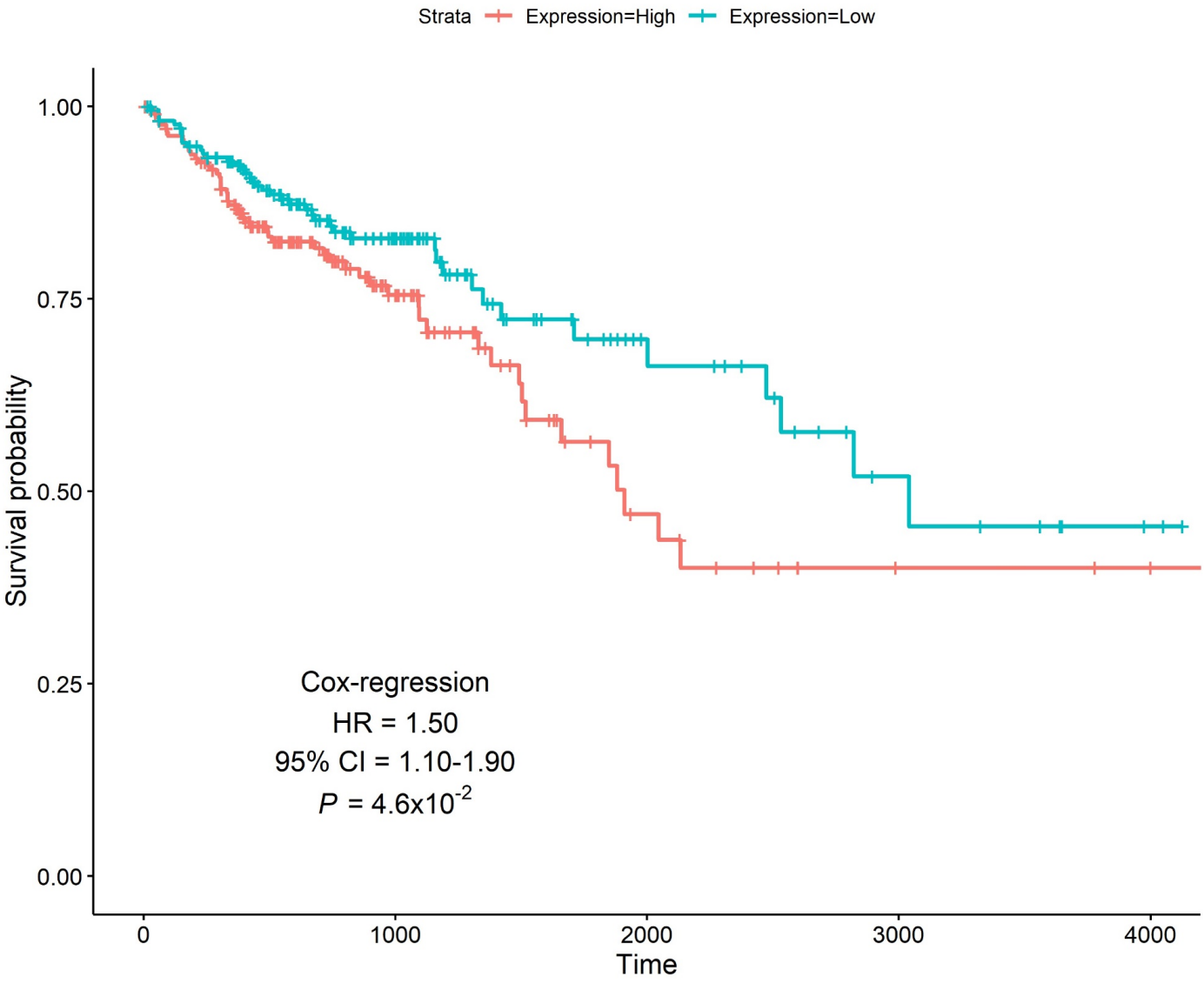


Figure 3.8. Kaplan-Meier plot for *ERBB4* expression levels in tumours from 438 patients with colon adenocarcinomas from the Human Protein Atlas. Time in days plotted against survival probability. High expression levels defined as median number of fragments per kilobase of exon per million reads >0. Cox-regression used to calculate *P*-value for differences in survival between the groups.

3.4 Discussion

3.4.1 No observed pleiotropic effects on survival

Despite identifying 18 somatic and clinicopathological factors that significantly influenced survival in COIN and COIN-B, we found that SNPs associated with these factors did not themselves affect survival thereby excluding potential pleiotropic effects. To generate a comprehensive genome-wide analysis of survival, we included prognostic factors into our multivariate analyses and observed little genomic inflation supporting the validity of this approach. rs142358223, which showed a significant association with white blood cell count, was not identified by Astle et al. (2016) in their analysis of human blood cell trait variation in the UK Biobank and INTERVAL studies, nor was any SNPs in strong LD with rs142358223.

3.4.2 Variation in *ERBB4* may predict survival in advanced CRC

The most significant SNP identified was rs79612564 which lies within intron 3 of *ERBB4*, a member of the EGFR subfamily. We confirmed the quality of the genotyping and imputation for this SNP via an independent assay. Patients carrying the minor allele had an additive effect on survival with a median decrease in life expectancy of approximately 40 days per allele carried in the advanced disease setting. rs79612564 was also significant in stage IV patients from SOCCS and, combined with COIN and COIN-B, reached genome wide significance. Our genetic data was supported by mechanistic data for this gene and we found that patients with high *ERBB4* expression in their colon

adenocarcinomas had worse survival. Furthermore, it has previously been shown that *ERBB4* over-expression in experimental systems enhances the survival and growth of cells driven by *Ras* and/or *Wnt* signalling (Williams et al. 2015).

However, rs79612564 was not replicated in stage IV patients from ISACC, nor in all patients from SOCCS and ISACC combined. This warrants further investigation, although it is noteworthy that overexpression and heterodimerization of ERBB4 and ERBB2 shows a significant association with late stage colorectal carcinomas (Lee et al. 2002). Therefore, it is possible that the association for rs79612564 can only be seen in patients with later stages of disease and survival in these patients is confounded by numerous clinical and pathological prognostic covariates which we accounted for in our GWAS but are, in general, not available in the population-based cohorts.

3.4.3 Potential clinical implications

In terms of clinical application, it should be noted that the effect size for rs79612564 is modest and will need to be combined with other prognostic factors to have any role in patient management. For example, our data suggests that this SNP acts independently of *KRAS* mutational status which itself is a prognostic factor. In isolation, rs79612564 has a HR of 1.24 but on a *KRAS* mutant background increases to 1.51. Although this effect size is still modest, it shows the potential for building germline, somatic and clinicopathological factors into a combined prognostic model.

3.4.4 Other independent loci

Most of the other loci of interest failed to be replicated or their directions of effect were opposite to those found in our discovery cohort. However, rs2050337 at 10q25.1 reached significance in the stage IV replication meta-analysis with a consistent direction of effect to COIN and COIN-B and was also significant in all patients from SOCCS. It lies approximately 500Kb upstream of *ADD3*, which encodes γ-Adducin, one subunit of Adducin; a ubiquitously expressed membrane-skeletal protein responsible for stabilization of the membrane cytoskeleton, cell signalling, ionic transportation, cell motility and cell-cell adhesion. *ADD3* has been associated with tumour growth and cell migration in breast (Yang et al. 2020), glioblastoma multiforme (Kiang et al. 2020) and lung cancer (Lechuga et al. 2019). In CRC, *ADD3* and its splicing isoform *ADD3-Ib* show decreased expression compared with normal mucosa, possibly contributing to the tissue's invasion ability (Luo and Shen 2017). However, even combined with COIN and COIN-B, rs2050337 still did not achieve genome-wide significance in patients with stage IV disease, suggesting that its effects, if genuine, are modest.

3.4.5 Power considerations and further study

Despite having 1,926 patients with advanced CRC (with a 75% event rate) in our GWAS, we lacked sufficient power to detect common alleles with low effect sizes (HR<1.3) at genome wide significance levels. Even by considering loci at suggestive significance levels, as we have done, we only had 33% power to detect common alleles with HRs of 1.2. Future studies will therefore have to combine their datasets for meta-analyses to

provide sufficient power to identify low impact alleles for survival. For example, to achieve 80% power to detect alleles with HRs of 1.2 and 1.1 would require 4,907 and 18,022 patients with a 75% event rate, respectively.

Chapter 4: Relationship between inherited genetic variation and survival from colorectal cancer stratified by tumour location

4.1 Introduction

4.1.1 Pathobiology of proximal, distal and rectal CRCs

Proximal and distal colonic cancers have distinct clinicopathological and molecular features, reflective of their embryological origin (Chapter 1, Section 1.1.4.1) and biology (Missiaglia et al. 2014) (Iacopetta 2002). Proximal colonic cancers are frequently *KRAS* (Rosty et al. 2013; Li et al. 2015b) and *BRAF* (Missiaglia et al. 2014; Li et al. 2015b) mutated, have MSI and a CpG island methylator phenotype (Sanz-Pamplona et al. 2011). They are more common in women and older patients, and while having a poorer prognosis, tend have a better response to 5FU chemotherapy (Iacopetta 2002). Distal cancers are typified by chromosomal abnormalities and aneuploidy (Bufill 1990). Rectal cancers have higher rates of locoregional relapse, a preference for lung metastases and a lower frequency of *KRAS* and *BRAF* mutations (Meguid et al. 2007; Phipps et al. 2013; Yang et al. 2016).

4.1.2 This study

The prognosis for patients with the same stage of CRC can vary and, in addition to clinicopathological features and somatic mutations, it is being recognised that germline variation also influences outcome. In Chapter 3, I identified germline variants

associated with survival in patients with advanced CRC from COIN and COIN-B. Given the inherent differences in the pathobiology of proximal and distal cancers, here I report on the impact of germline variation on CRC prognosis by tumour anatomical site.

4.2 Materials and methods

4.2.1 Patients and genotyping

1,948 patients from COIN and COIN-B had germline genotyping and survival data available. The minimum MAF for SNPs was set at 5%, leaving 2.9 million SNPs for analysis. See Chapter 2, Section 3.1 for full details on patients, DNA extraction, genotyping and QC.

I assigned patients to groups by location of their primary tumour (Labadie et al. 2022). Proximal tumours - those within the hepatic flexure, transverse colon, cecum and ascending colon (514 patients, 413 with events); Distal tumours - those within the descending colon, sigmoid colon and splenic flexure (n=493 patients, 358 with events); Rectal tumours - those within the rectosigmoid junction and rectum (892 patients with 645 events) (**Figure 4.1**). For 49 patients, data on primary tumour location was missing.

4.2.2 Replication cohort

To replicate findings, I used UKB patient data (Chapter 2, Section 2.3.4). CRC patients were stratified according to the location of their tumour - 1,433 (473 with events) with proximal disease, 1,450 (420 events) with distal disease and 1,869 (495 events) with rectal disease (**Figure 4.1**). For 326 patients there was insufficient information to assign the anatomical site of the CRC.

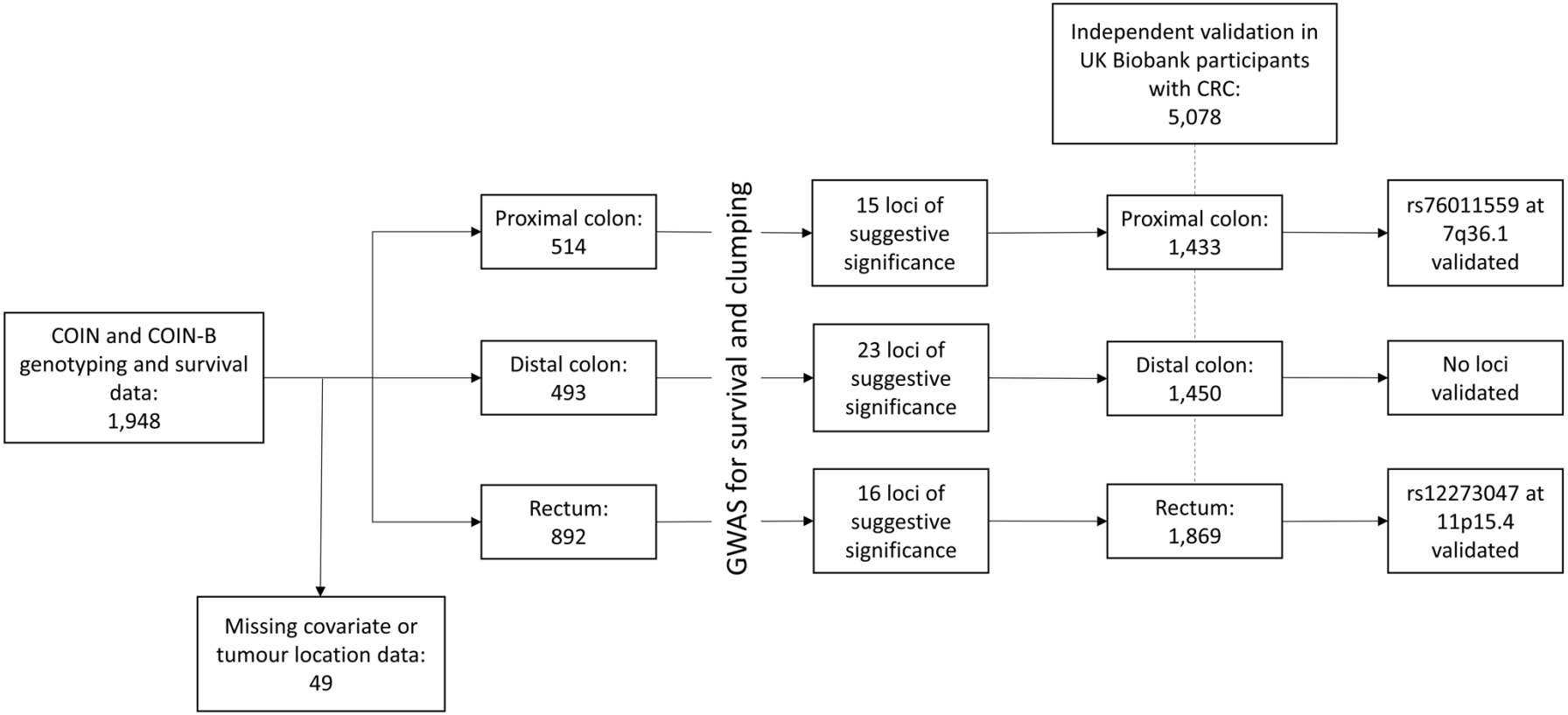


Figure 4.1. Flow diagram depicting the genetic and survival analyses of patients from COIN and COIN-B by primary tumour location. 514 patients had primary tumours in the proximal colon, 493 in the distal colon and 892 in the rectum. Lead SNPs from independent loci suggestive of association with survival were tested for replication in participants from the UK Biobank with proximal colon (n=1,433), distal colon (n=1,450) and rectal cancers (n=1,869), respectively.

4.2.3 Statistical analyses

I previously identified clinicopathological factors associated with survival (n=11) in patients from COIN and COIN-B (Chapter 3, Section 3.3.1). Dimensionality reduction was performed using PCA to reduce the risk of overfitting (Chapter 2, Section 2.4.2) - the first five were selected, explaining 78-80% of the total variance (**Figure 2.1**). I carried out GWAS for OS by location of the primary tumour under an additive model. For any SNPs suggestive of an association, I performed clumping and tested the lead SNPs at each independent locus (n=54) in replication cohorts from the UKB. $P<0.05$ was used as the significance threshold for replication.

Power to detect an effect of rs313566 on survival in UKB patients with proximal, distal, and rectal tumours was estimated using an additive model, HR=0.52 (observed in COIN and COIN-B), $P=0.05$ and sample sizes of 1433 (473 events), 1450 (420 events) and 1869 (495 events), respectively.

To increase the power to detect associations, I also performed GWAS for survival in UKB patients by location of their colorectal tumour, using age and sex as covariates, followed by genome-wide meta-analysis with the COIN and COIN-B data using a fixed-effects model implemented in PLINK v1.9.

Gene and gene-set analysis was completed on the summary statistics from the association analysis to identify genes containing significant numbers of highly associated SNPs and significantly enriched gene-sets (Chapter 2, Section 2.4.5).

4.2.4 Bioinformatic analyses

See Chapter 2, Sections 2.4.3, 2.5.1 and 2.3.5 for details on GWAS analysis, LocusZoom plots and eQTL analyses, respectively.

I sought an association between Phosphatidylinositol 4-Kinase Type 2 Beta (*PI4K2B*) expression levels in colorectal tumours and survival in 597 CRC patients from THPA (Chapter 2, Section 2.3.7). Samples were classified as high expression using a threshold of FPKM>7.38 as per THPA recommendations. I also performed survival analysis using a linear Cox proportional-hazards model.

4.3 Results

4.3.1 Clinicopathological features of patients stratified by tumour location

1,899 patients from COIN and COIN-B had genotyping, survival, clinicopathological and primary tumour location data available (**Figure 4.1**). Patients with proximal CRC (n=514) had a higher frequency of *KRAS* (39.1%) and *BRAF* (16.0%) mutations and worse prognosis (median survival 397 days) compared to patients with distal CRC (n=493, 25.6%, 4.3% and 514 days, all $P<1.0x10^{-4}$, respectively) and rectal cancers (n=892, 33.3%, $P=1.2x10^{-2}$; 4.1%, $P<1.0x10^{-4}$ and 520 days, $P<1.0x10^{-4}$, respectively) (**Table 4.1**).

4.3.2 Relationship between germline variation and survival by tumour location

Genome-wide survival analyses of patients from COIN and COIN-B were stratified by primary tumour location. There was no detectable genomic inflation (λ=1.03-1.12). No SNPs passed genome-wide significance regardless of tumour location (**Figure 4.2**).

SNPs at 15 independent loci were suggestive of an association with survival in patients with tumours in the proximal colon, 23 loci in those with tumours in the distal colon and 16 loci in those with tumours in the rectum (**Figure 4.2**, **Table 4.2**). I sought independent replication of lead SNPs at each of these loci in 5,078 UKB participants. rs76011559 mapping to 7q36.1 (123kb upstream of *CUL1*) replicated in patients with proximal tumours (HR=1.31, 95% CI=1.03-1.66, $P=2.8x10^{-2}$, **Figure 4.3**, **Table 4.2**). In the advanced disease setting, patients carrying at least one copy of the minor (C)

allele had a median reduction in survival of 121 days compared to patients homozygous for the major (A) allele (**Figure 4.3**).

rs12273047 at 11p15.4 replicated in patients with rectal tumours (HR=1.19, 95% CI=1.03-1.38, P=1.6x10^{-2}; **Figure 4.4**, **Table 4.2**). Patients carrying at least one copy of the minor (C) allele had a median reduction in survival of 132 days compared to patients homozygous for the major (T) allele (**Figure 4.4)**. No other lead SNPs were replicated (**Table 4.2**).

Clinicopathological factor		Proximal tumour (n = 514)		Distal tumour (n = 493)		Rectum (n = 892)		*P*
		n	%	n	%	n	%	
Sex	Male	307	59.7	312	63.3	625	70.1	$2.2x10^{-4}$
	Female	207	40.3	181	36.7	267	29.9	
Age	Median (years)	65	-	64	-	63	-	-
Overall survival	Median days (95% CI)	397 (359-444)	-	514 (471-556)	-	520 (496-581)	-	$<1.0x10^{-4}$
WHO performance status	0	216	42.0	209	42.4	459	51.5	$1.3x10^{-3}$
	1	251	48.8	249	50.5	375	42.0	
	2	47	9.1	35	7.1	58	6.5	
Status of primary tumour	Resected	316	61.5	270	54.8	421	47.2	$<1.0x10^{-4}$
	Unresected	198	38.5	223	45.2	471	52.8	
Timing of metastases	Metachronous	136	26.5	119	24.1	311	34.9	$<1.0x10^{-4}$
	Synchronous	378	73.5	374	75.9	581	65.1	
Type of metastases	Liver only	86	16.7	151	30.6	185	20.8	$<1.0x10^{-4}$
	Liver plus others	272	52.9	255	51.7	474	53.3	
	Non-liver	156	30.4	87	17.6	231	26.0	
Number of metastatic sites	0	0	0.0	0	0.0	2	0.2	0.23
	1	175	34.0	196	39.8	310	34.8	
	2	200	38.9	181	36.7	367	41.1	
	≥3	139	27.0	116	23.5	213	23.9	
KRAS status	Mutated	201	39.1	126	25.6	297	33.3	$<1.0x10^{-4}$
	Wild-type	224	43.6	283	57.4	453	50.8	
	n/k	89	17.3	84	17.0	142	15.9	
NRAS status	Mutated	16	3.1	20	4.1	30	3.4	0.66
	Wild-type	397	77.2	373	75.7	699	78.4	
	n/k	101	19.6	100	20.3	163	18.3	
BRAF status	Mutated	82	16.0	21	4.3	37	4.1	$<1.0x10^{-4}$
	Wild-type	332	64.6	373	75.7	695	77.9	
	n/k	100	19.5	99	20.1	160	17.9	
PIK3CA status	Mutated	62	12.1	45	9.1	79	8.9	0.065
	Wild-type	308	59.9	315	63.9	594	66.6	
	n/k	144	28.0	133	27.0	219	24.6	

Table 4.1. Clinicopathological features of COIN and COIN-B patients by tumour site. Data are n (%) or median. Differences between patients were analysed using a

Chi-squared test, Fisher's exact test (for number of metastatic sites) or log rank test (for overall survival). *Non-liver metastases included those in the lungs, peritoneum and lymph nodes. n/k – not known - some data for somatic mutation status was not known due to the lack of availability of tumour tissue or failed amplification.

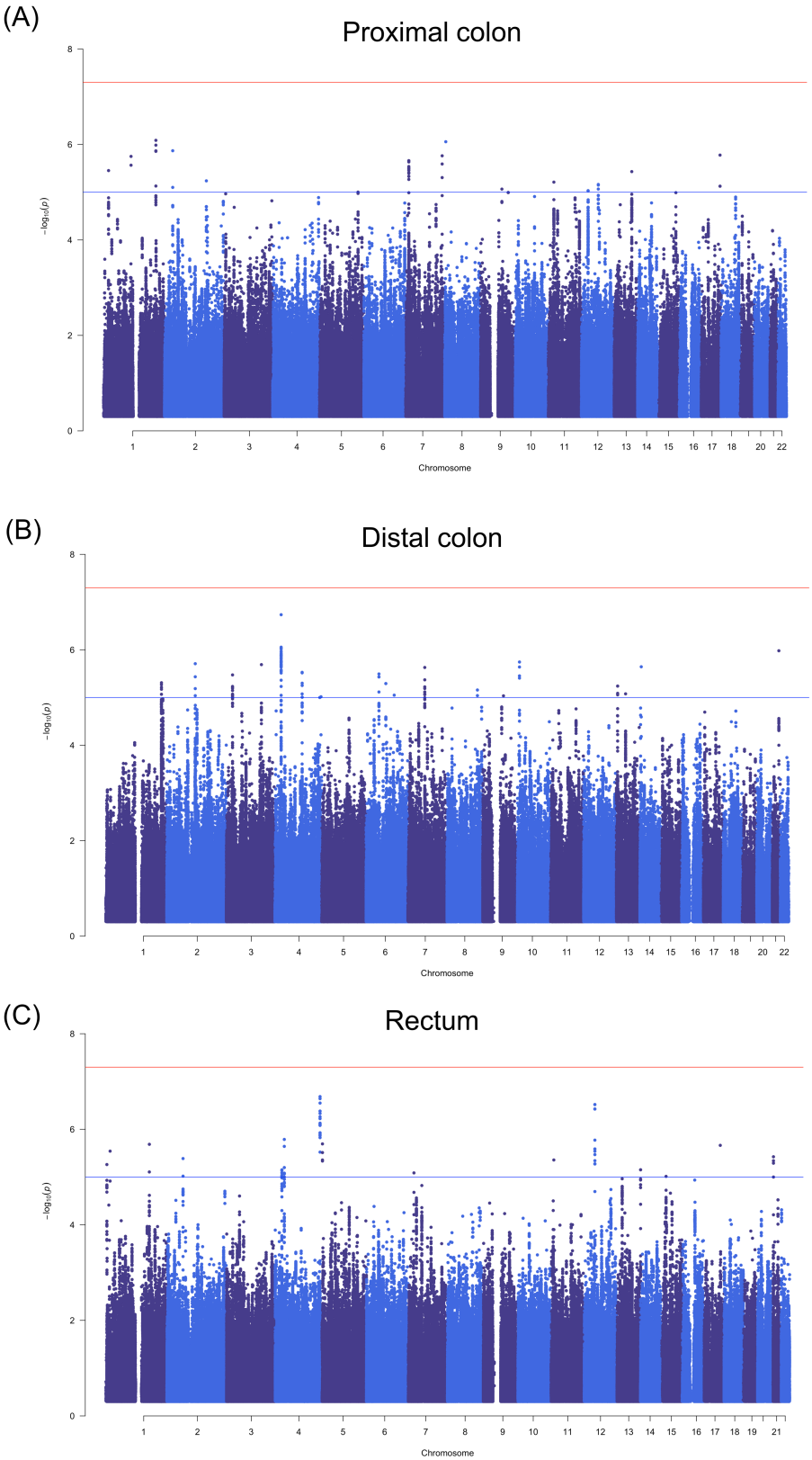


Figure 4.2. Manhattan plots of single nucleotide polymorphism (SNP) associations with overall survival (OS) in patients from COIN and COIN-B with primary tumours in (A) the proximal colon (n=514), (B) the distal colon (n=493) and (C) the rectum (n=892). SNPs are ordered by chromosome position and plotted against the $-\log_{10}(P)$ for their association with OS. The red line represents the threshold for genome-wide significance ($P<5.0\times10^{-8}$) and the blue line is the threshold for suggestive significance ($P<1.0\times10^{-5}$).

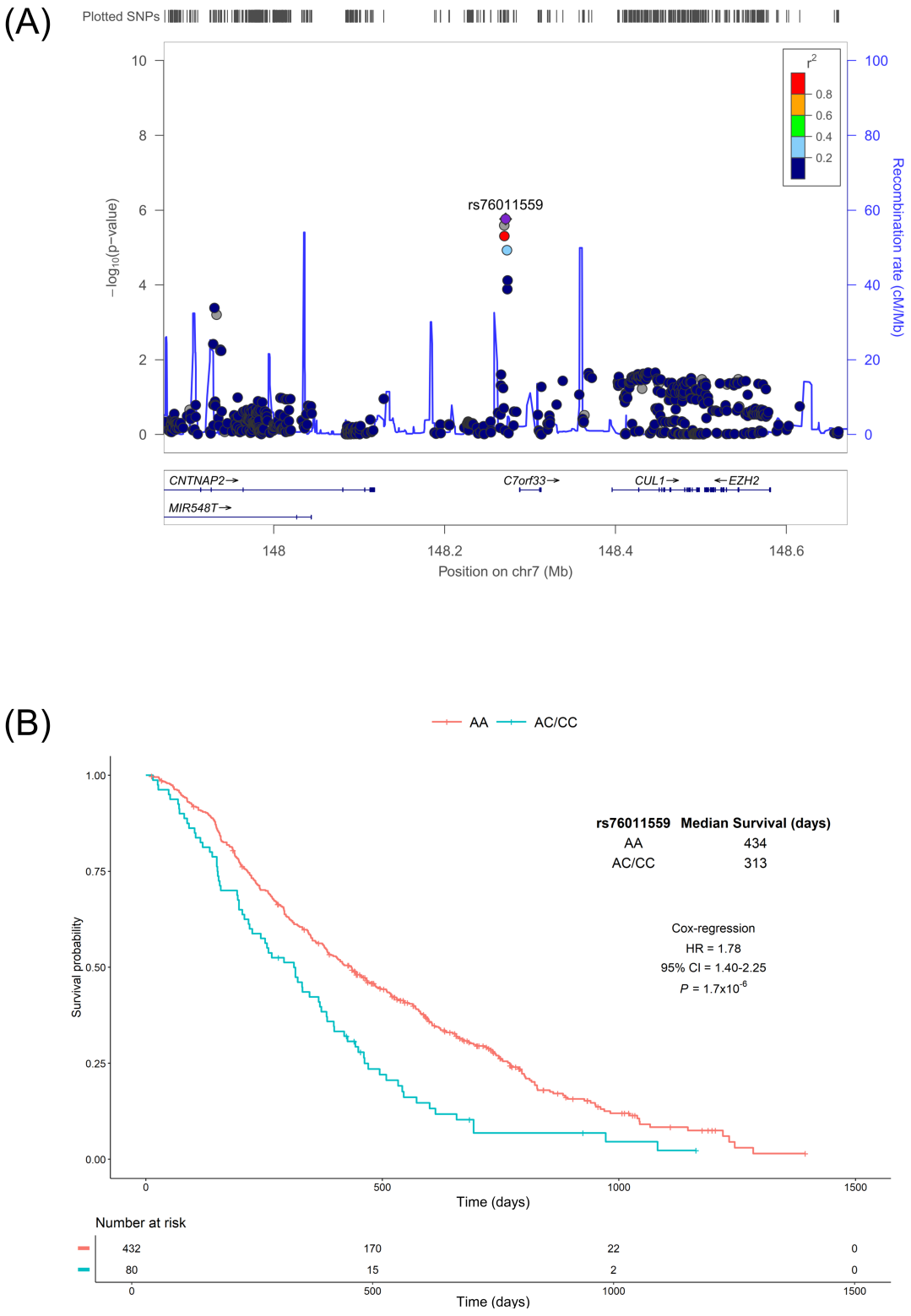


Figure 4.3. Relationship between rs76011559 genotype and overall survival (OS) in patients from COIN and COIN-B with proximal colon tumours. (A) Regional locus zoom plot shows results of the analysis for single nucleotide polymorphisms (SNPs) and recombination rates. $-\log_{10}(P)$ (y axis) of the SNPs are

shown according to their chromosomal positions (x axis) for an area 400Kb upstream and downstream of rs76011559 (in purple). The colour intensity of each symbol reflects the extent of linkage disequilibrium with the sentinel SNP, deep blue (r^2=0) through to dark red (r^2=1.0). Genetic recombination rates, estimated using 1000 Genomes Project samples, are shown with a blue line. Physical positions are based on NCBI build 37 of the human genome. Also shown are the relative positions of genes and transcripts mapping to the region of association. Genes have been redrawn to show their relative positions; therefore, maps are not to physical scale. **(B)** Kaplan-Meier plot of the relationship between rs76011559 genotype and OS. Time in days plotted against survival probability for patients homozygous for the major allele (AA) and heterozygous (AC) or homozygous for the minor allele (CC). The number of patients still at risk at each time point is shown beneath.

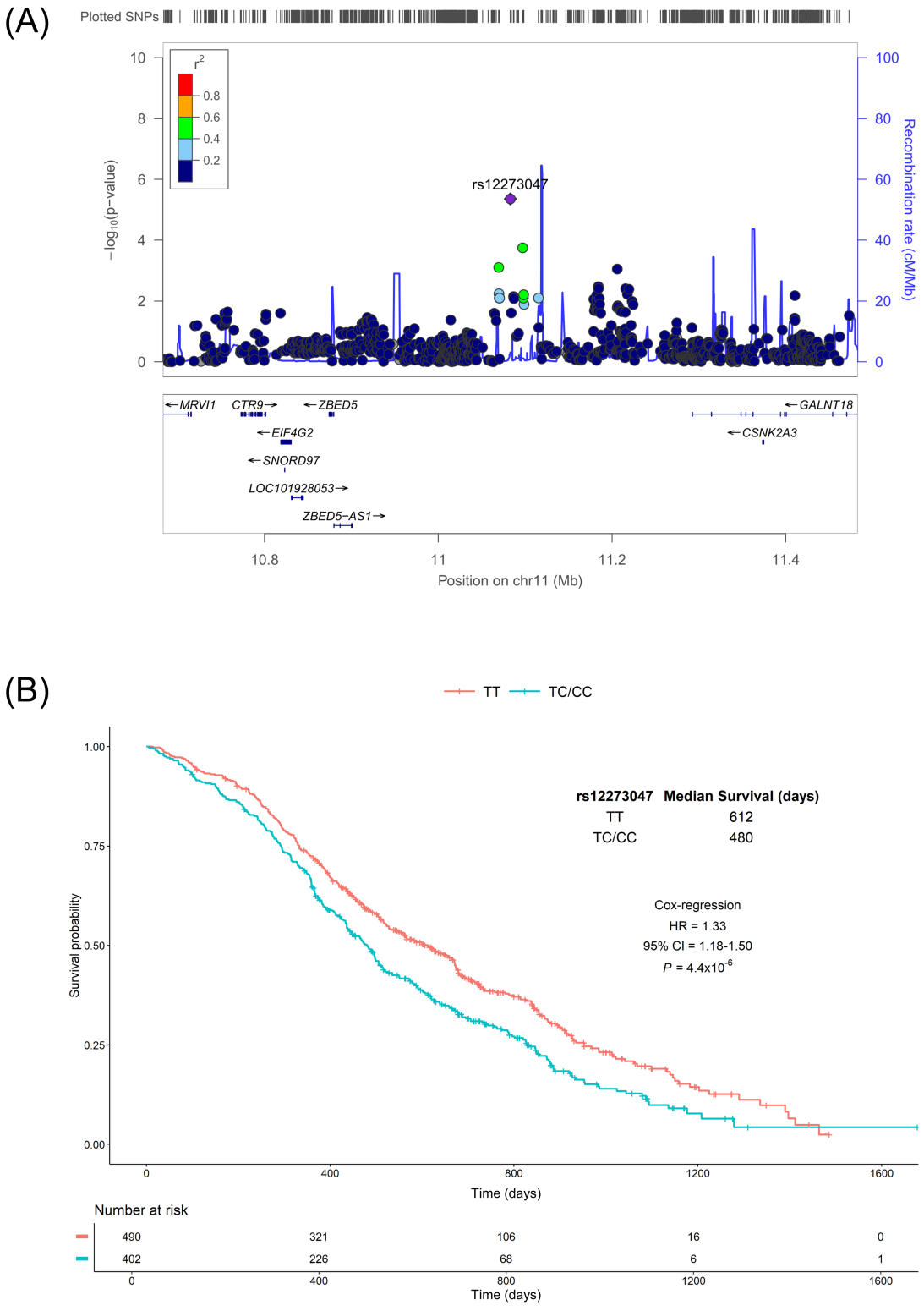


Figure 4.4. Relationship between rs12273047 genotype and overall survival in patients from COIN and COIN-B with rectal tumours. (A) Regional locus zoom plot shows results of the analysis for single nucleotide polymorphisms (SNPs) and

recombination rates. $-\log_{10}(P)$ (y axis) of the SNPs are shown according to their chromosomal positions (x axis) for an area 400Kb upstream and downstream of rs12273047 (in purple). The colour intensity of each symbol reflects the extent of linkage disequilibrium with the sentinel SNP, deep blue (r^2=0) through to dark red (r^2=1.0). Genetic recombination rates, estimated using 1000 Genomes Project samples, are shown with a blue line. Physical positions are based on NCBI build 37 of the human genome. Also shown are the relative positions of genes and transcripts mapping to the region of association. Genes have been redrawn to show their relative positions; therefore, maps are not to physical scale. **(B)** Kaplan-Meier plot of the relationship between rs12273047 genotype and overall survival. Time in days plotted against survival probability for patients homozygous for the major allele (TT) and heterozygous (TC) or homozygous for the minor allele (CC). The number of patients still at risk at each time point is shown beneath.

Primary tumour location	SNP	Locus	Minor allele	Genes	COIN and COIN-B			UK Biobank		
					HR	95% CI	*P*	HR	95% CI	*P*
	rs12062055	1q32.3	G		2.02	1.53-2.67	8.2×10^{-7}	0.90	0.68-1.19	0.46
	rs4304342	8p23.2	C	*CSMD1*	0.67	0.57-0.79	8.8×10^{-7}	0.98	0.84-1.13	0.77
	rs62135742	2p22.3	C	*LTBP1*	1.80	1.42-2.29	1.4×10^{-6}	0.97	0.78-1.20	0.75
	rs147899046*	17q25.3	A	*DNAH17*	1.43	1.23-1.65	1.7×10^{-6}	1.11	0.97-1.27	0.14
	rs76011559	7q36.1	C		1.78	1.40-2.25	1.7×10^{-6}	1.31	1.03-1.66	**2.8×10^{-2}**
	rs10857917	1p13.2	G	*LOC643355*	1.44	1.24-1.67	1.8×10^{-6}	0.97	0.84-1.12	0.67
	rs6460936	7p21.3	C	*TMEM106B, VWDE*	1.57	1.30-1.90	2.2×10^{-6}	1.00	0.83-1.21	0.99
	rs35955655*	1p36.12	CTA	*CDA, DDOST, MIR6084, PINK1, PINK1-AS*	0.71	0.62-0.82	3.5×10^{-6}	1.05	0.92-1.19	0.47
Proximal	rs1388194	13q31.3	T		0.71	0.62-0.82	3.7×10^{-6}	0.93	0.81-1.06	0.29
	rs112651521	2q31.1	T	*BBS5, FASTKD1, KLHL41, PPIG*	1.71	1.36-2.16	5.8×10^{-6}	0.99	0.80-1.24	0.96
	rs1514081	11p14.3	C		0.73	0.63-0.83	6.1×10^{-6}	0.97	0.85-1.10	0.64
	rs10878838	12q15	T	*LOC100507195*	1.64	1.32-2.03	6.9×10^{-6}	1.09	0.88-1.35	0.44
	rs148684057	9q21.32	GT	*LOC101927575*	1.72	1.35-2.19	8.6×10^{-6}	0.98	0.80-1.30	0.89
	rs11048907	12p11.23	T	*ARNTL2, C12orf71, MED21, STK38L, TM7SF3*	1.71	1.35-2.16	9.3×10^{-6}	1.06	0.86-1.32	0.57
	rs78738433	5q33.3	C	*ADAM19, CYFIP2, NIPAL4*	1.90	1.43-2.52	1.0×10^{-5}	1.04	0.81-1.33	0.77
	rs313566	4p15.2	A	*ANAPC4, PI4K2B, SEPSECS, SEPSECS-AS1, ZCCHC4*	0.52	0.41-0.67	1.8×10^{-7}	1.15	0.93-1.42	0.19
	rs2837637*	21q22.2	A	*DSCAM*	1.47	1.26-1.72	1.0×10^{-6}	1.10	0.96-1.26	0.17

	rs7907707	10p14	C		1.63	1.33-1.99	$1.8x10^{-6}$	1.04	0.86-1.26	0.70
	rs10182527	2q14.1	T	*DPP10, DPP10-AS1*	1.44	1.24-1.67	$2.0x10^{-6}$	1.08	0.95-1.24	0.24
	rs76041099	3q23	C	*LOC100507389*	2.14	1.57-2.94	$2.0x10^{-6}$	0.83	0.61-1.14	0.26
	rs11159167	14q12	G		1.43	1.23-1.67	$2.3x10^{-6}$	0.97	0.84-1.12	0.69
	rs117589090	10p14	G		2.08	1.53-2.81	$2.3x10^{-6}$	0.89	0.64-0.24	0.50
	rs4718825	7q11.22	G		1.55	1.29-1.87	$2.3x10^{-6}$	0.98	0.82-1.17	0.83
	rs7656285	4q25	C	*LRIT3, RRH*	1.42	1.22-1.64	$3.0x10^{-6}$	0.93	0.81-1.07	0.34
	rs6921841	6p12.2	A		1.62	1.32-1.98	$3.2x10^{-6}$	1.05	0.88-1.26	0.56
	rs10510552	3p24.2	T		1.45	1.24-1.69	$3.4x10^{-6}$	0.88	0.76-1.00	0.06
Distal	rs34507557	1q42.13	CT	*CDC42BPA*	1.66	1.34-2.07	$4.9x10^{-6}$	1.10	0.91-1.34	0.33
	rs28583014	4q25	A	*EGF, ELOVL6*	1.73	1.37-2.20	$5.0x10^{-6}$	0.93	0.74-1.17	0.53
	rs2057331	6q14.1	G	*C6orf7*	1.80	1.40-2.33	$5.1x10^{-6}$	0.96	0.75-1.23	0.75
	rs41268739	1q42.13	T	*CDC42BPA*	2.04	1.50-2.78	$5.4x10^{-6}$	0.90	0.65-1.25	0.54
	rs9995789	4q25	T	*ELOVL6*	1.52	1.27-1.83	$5.6x10^{-6}$	0.98	0.82-1.17	0.84
	rs7319699	13q12.12	G	*TNFRSF19*	1.45	1.24-1.71	$5.8x10^{-6}$	1.10	0.95-1.27	0.21
	rs7826050	8q24.13	G	*DERL1*	1.45	1.23-1.70	$7.0x10^{-6}$	0.99	0.85-1.15	0.87
	rs11842682	13q21.1	T		1.51	1.26-1.81	$8.4x10^{-6}$	0.94	0.79-1.12	0.50
	rs1033393	6q22.1	T		1.57	1.29-1.92	$8.9x10^{-6}$	1.02	0.85-1.23	0.80
	rs2796466	9q21.32	T	*TLE1*	1.41	1.21-1.64	$9.2x10^{-6}$	0.87	0.76-1.01	0.06
	rs7660386	4q35.2	G		0.66	0.55-0.79	$9.6x10^{-6}$	0.95	0.81-1.11	0.51
	rs72702433	4q34.3	G		1.86	1.41-2.44	$1.0x10^{-5}$	0.97	0.74-1.30	0.87
	rs73011737	4q34.3	T		1.68	1.38-2.04	$2.1x10^{-7}$	0.97	0.78-1.22	0.82
	rs77984832	12q12	T		1.82	1.45-2.29	$3.0x10^{-7}$	0.87	0.67-1.12	0.28
	rs1562098	4p14	T		1.32	1.18-1.48	$1.6x10^{-6}$	0.99	0.86-1.13	0.85
	rs10067149	5p15.33	G		1.31	1.17-1.47	$2.0x10^{-6}$	1.04	0.92-1.19	0.50
	rs74602176	1q25.2	A	*BRINP2*	1.72	1.38-2.15	$2.1x10^{-6}$	0.91	0.69-1.21	0.53
	rs2949938	17q24.2	A	*PITPNC1*	1.69	1.36-2.10	$2.2x10^{-6}$	0.98	0.71-1.34	0.90
	rs60453441	1p36.13	G		0.69	0.59-0.81	$2.9x10^{-6}$	1.02	0.87-1.20	0.81
Rectal	rs2822995	21q11.2	T	*NRIP1*	1.37	1.20-1.56	$3.8x10^{-6}$	1.13	0.97-1.33	0.12

rs268872	2p14	T	*ACTR2*	1.39	1.21-1.60	$4.1x10^{-6}$	0.98	0.84-1.15	0.81
rs12273047	11p15.4	C		1.33	1.18-1.50	$4.4x10^{-6}$	1.19	1.03-1.38	**$1.6x10^{-2}$**
rs35066664	1p36.32	G		1.69	1.35-2.11	$5.5x10^{-6}$	0.98	0.77-1.27	0.90
rs34529111	4p14	G		1.45	1.24-1.71	$6.3x10^{-6}$	1.09	0.90-1.31	0.37
rs112063020	13q34	AGTTT	*CDC16, UPF3A*	1.31	1.17-1.48	$7.0x10^{-6}$	1.07	0.93-1.23	0.36
rs16878917	4p15.2	A		0.74	0.64-0.84	$7.1x10^{-6}$	0.98	0.84-1.14	0.78
rs113230287	7p15.3	C	*STEAP1B*	1.45	1.23-1.72	$8.2x10^{-6}$	0.86	0.71-1.05	0.14
rs78745358	15q14	A	*C15orf41*	1.63	1.31-2.02	$9.7x10^{-6}$	1.01	0.72-1.34	0.93

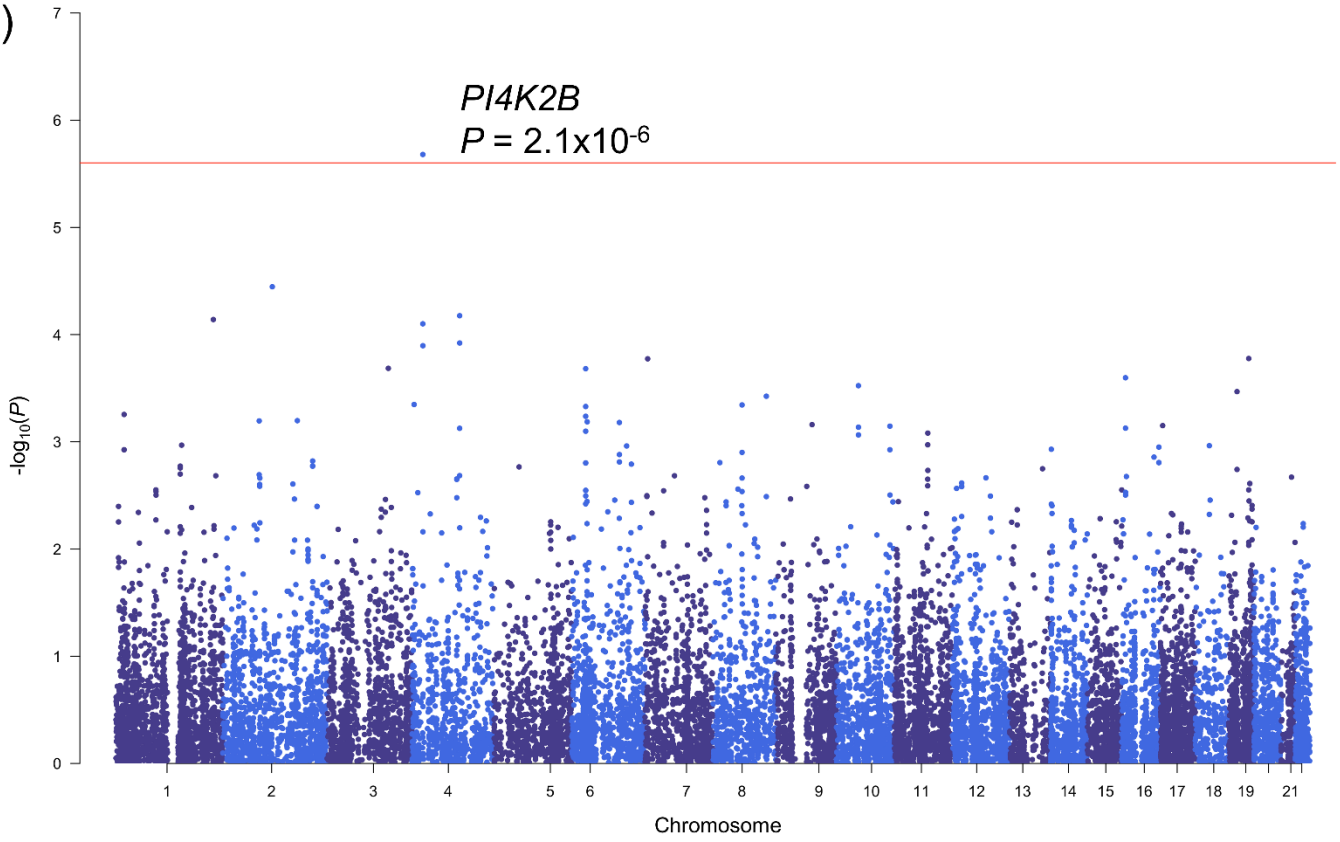
Table 4.2. Replication of loci suggestive of association with survival in COIN and COIN-B. Independent replication of lead single nucleotide polymorphisms was carried out using participants from the UK Biobank (UKB) with proximal colon, distal colon and rectal tumours. Tumour location, SNP location, minor allele, overlapping genes, Hazard Ratio, 95% confidence intervals and *P*-value are listed for survival (time from trial recruitment to death or end of study for COIN and COIN-B, and time from diagnosis to death or data distribution date for the UKB). rs76011559 replicated in patients with proximal tumours and rs12273047 replicated in patients with rectal tumours (highlighted in bold). *rs35955655, rs147899046 and rs2837637 were not available in the UKB and so were replaced with the proxies rs12021613 (1000 genomes project R^2=1 and D'=1), rs4969218 (r^2=0.99 and D'=1) and rs1012846 (r^2=0.6 and D'=1), respectively.

4.3.3 Gene and expression analyses

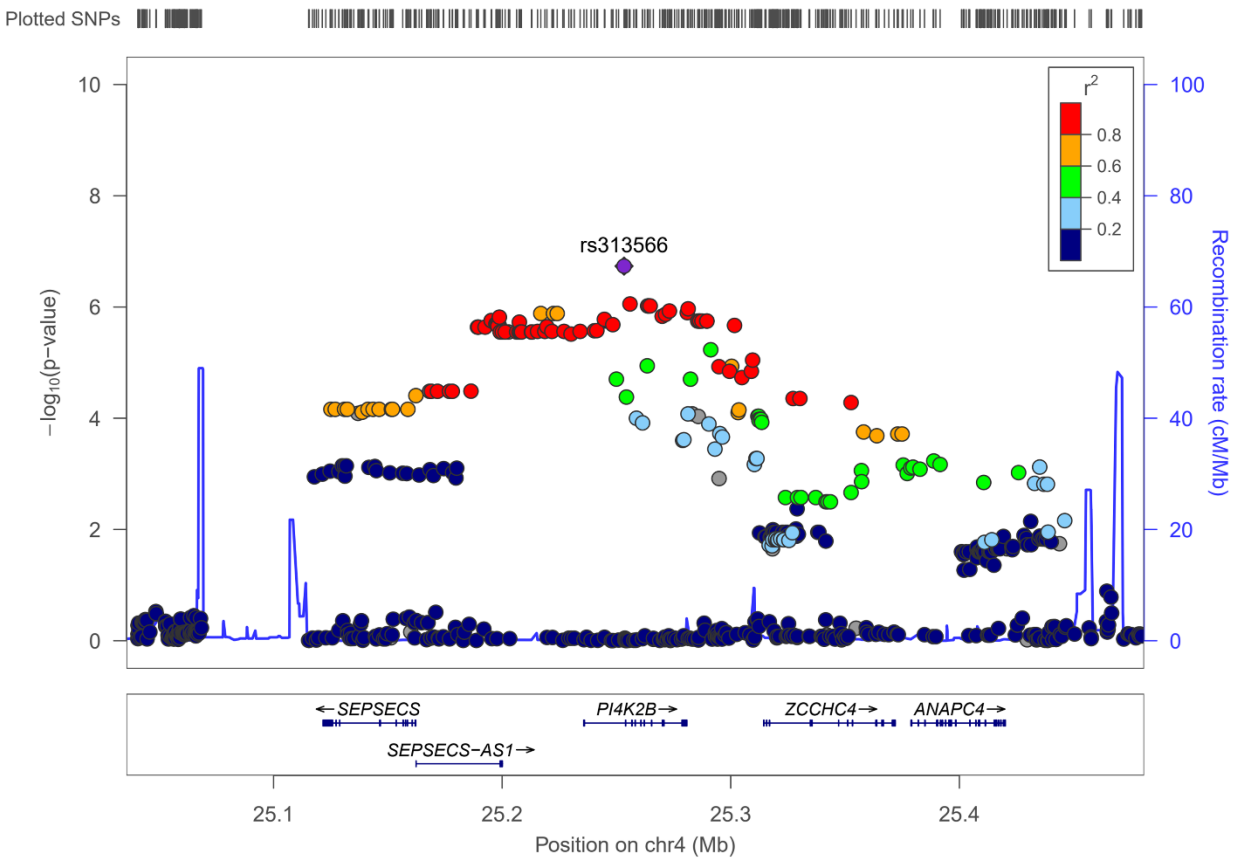
In MAGMA gene analyses, only *PI4K2B* was significantly associated with survival in COIN and COIN-B patients with distal cancers, beyond the threshold for multiple testing (P=2.1x10^{-6}; **Figure 4.5**). Patients carrying one copy of the minor (A) allele in the lead SNP, rs313566 in intron 1 of *PI4K2B*, had a median increase in survival of 245 days compared to patients homozygous for the major (G) allele (HR=0.52, 95% CI=0.4-0.7, P=1.8x10^{-7}, **Figure 4.5**). In contrast, rs313566 genotype was not associated with survival in patients with proximal cancers (HR=1.10, 95% CI=0.89-1.36, P=0.37, $P_{Z\text{-}test}$ compared to distal cancers=6.5x10^{-6}) or those with rectal cancers (HR=1.16, 95% CI=0.97-1.39, P=0.09, $P_{Z\text{-}test}$ compared to distal cancers=1.9x10^{-7}).

I sought further mechanistic understanding of rs313566. rs313566 was an eQTL for *PI4K2B* in several cell types (cultured fibroblasts, cerebellum, cerebellar hemisphere, sun exposed skin, tibial nerve, and spleen; P<3.8x10^{-5}) with the A-allele associated with increased *PI4K2B* expression. I found that higher *PI4K2B* expression in tumour tissue was associated with improved survival in 597 unrelated patients with colorectal tumours from THPA (log rank P=9.6x10^{-5}, **Figure 4.6**). This finding was replicated under a linear Cox-proportional hazards model (HR=0.94, 95% CI=0.9-1.0, P=7.0x10^{-3}). Despite this, I failed to replicate the association between rs313566 and survival in UKB patients with distal (HR=1.15, 95% CI=0.93-1.42, P=0.19), proximal (HR=1.03, 95% CI=0.84-1.29, P=0.74) or rectal (HR=1.11, 95% CI=0.91-1.34, P=0.29) cancers, despite having over 99% power.

(A)


PI4K2B
P = 2.1x10-6
-log10(P)
Chromosome

(B)


Plotted SNPs
rs313566
r2
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
SEPSECS
PI4K2B
ZCCHC4
ANAPC4
SEPSECS−AS1
Position on chr4 (Mb)

(C)

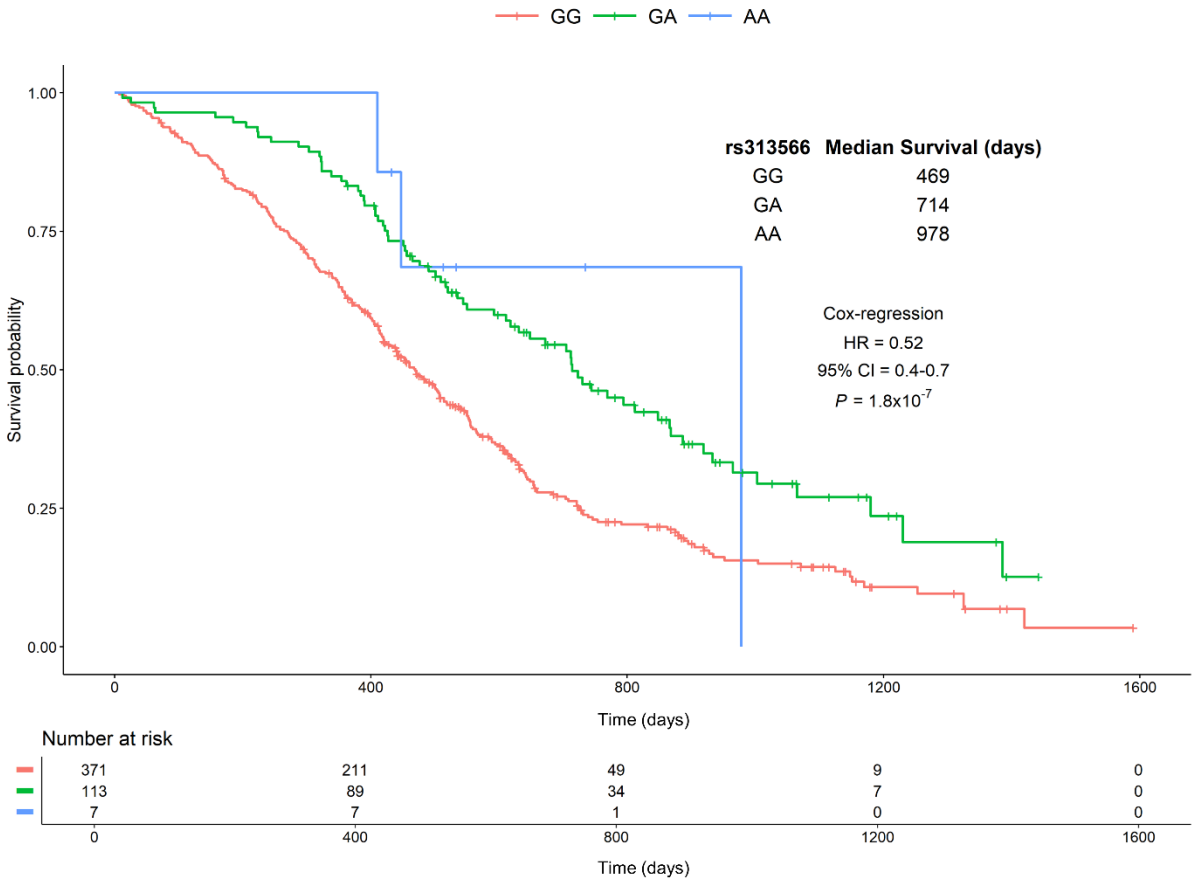


Figure 4.5. Relationship between gene, genotype and survival in patients from COIN and COIN-B with primary tumours in the distal colon. (A) Manhattan plot of gene associations with overall survival (OS). Genes are ordered by chromosome position and plotted against the -$\log_{10}$(*P*) for their association with OS. The red line represents the threshold for genome-wide significance (P=2.5x10^{-6}). **(B)** Regional locus zoom plot shows results of the analysis for single nucleotide polymorphisms (SNPs) and recombination rates. −$\log_{10}$(*P*) (y axis) of the SNPs are shown according to their chromosomal positions (x axis) for an area 200Kb upstream and downstream of *PI4K2B*. The sentinel SNP (purple) is labelled by its rsID (rs313566). The colour intensity of each symbol reflects the extent of linkage disequilibrium with the sentinel SNP, deep blue (r^2=0) through to dark red (r^2=1.0). Genetic recombination rates, estimated using 1000 Genomes Project samples, are shown with a blue line. Physical positions are based on NCBI build 37 of the human genome. Also shown are the relative positions of genes and transcripts mapping to the region of association. Genes have been redrawn to show their relative positions; therefore, maps are not to physical scale. **(C)** Kaplan-Meier plot of the relationship between rs313566 genotype and OS. Time in days plotted against survival probability for patients homozygous for the major allele (GG) and heterozygous (GA) or homozygous for the minor allele (AA). The number of patients still at risk at each time point is shown beneath.

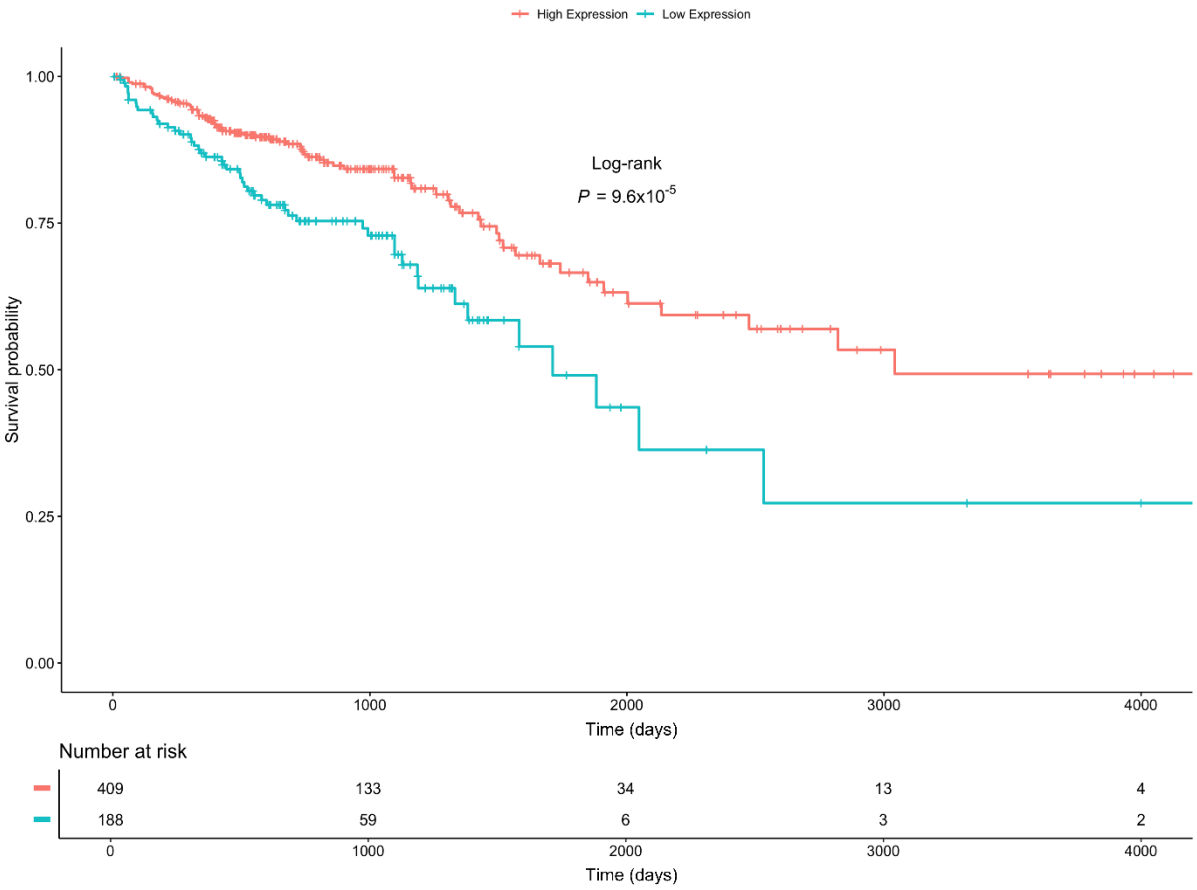


Figure 4.6. Kaplan-Meier plot for *PI4K2B* expression levels in colorectal tumours from 597 patients from the Human Protein Atlas. Time in days plotted against survival probability. High expression levels defined as median number of fragments per kilobase of exon per million reads >7.38. A log-rank test was used to calculate *P*-value for differences in survival between the groups.

4.3.4 Gene-set analyses

Four gene-sets (negative regulation of phospholipid biosynthetic process, phosphatidic acid biosynthetic process, 1-acylglycerophosphocholine O-acyltransferase activity and long-term memory) reached significance beyond multiple testing thresholds in patients from COIN and COIN-B with rectal cancers (**Table 4.3**).

Primary tumour location	GO Term	Gene-Set Name	*P*	*q*
Rectal	GO:0071072	Negative regulation of phospholipid biosynthetic process	$6.7x10^{-11}$	$6.6x10^{-7}$
	GO:0006654	Phosphatidic acid biosynthetic process	$5.6x10^{-7}$	$2.8x10^{-3}$
	GO:0047184	1-acylglycerophosphocholine O-acyltransferase activity	$8.5x10^{-6}$	$2.8x10^{-2}$
	GO:0007616	Long term memory	$1.6x10^{-5}$	$3.9x10^{-2}$

Table 4.3. MAGMA gene-set analysis for survival in patients from COIN and COIN-B by tumour location. Statistically significant sets with $q<0.05$ are presented. Gene-ontology (GO) term, full descriptive name, *P*-value and corrected *P*-value (*q*) are shown.

4.3.5 Meta-analysis of COIN, COIN-B and UKB by tumour location

To increase our power to detect associations, I carried out GWAS for survival in UKB patients by tumour location and meta-analysed the data with COIN and COIN-B. No SNPs reached genome-wide significance although three SNPs were close to this threshold in patients with rectal tumours (rs3980660 at 2q14.3, HR=0.79, 95% CI=0.61-0.97, $P=2.2x10^{-7}$; rs17237514 at 15q22.2, HR=0.73, 95% CI=0.50-0.97, $P=2.9x10^{-7}$ and rs12273047 at 11p15.4, HR=1.27, 95% CI=1.09-1.46, $P=4.1x10^{-7}$). No genes reached genome-wide significance. Three gene-sets reached significance in patients with rectal cancers (negative regulation of phospholipid biosynthetic process, $P=9.6x10^{-12}$, $q=9.5x10^{-8}$; phosphatidic acid biosynthetic process, $P=8.2x10^{-8}$,

q=4.1x10^{-4} and positive regulation of response to endoplasmic reticulum stress, P=1.4x10^{-5}, q=4.5x10^{-2}).

4.3.6 Relationship between previously reported prognostic SNPs and tumour location

Three SNPs have been associated with CRC survival by tumour location (Labadie et al. 2022). rs698022 was not replicated in patients from COIN and COIN-B despite having 84% power. rs189655236 also failed replication but with more limited power (54%). However, rs144717887 (INFO score=0.92) was replicated and associated with improved survival in patients with proximal tumours under multivariate analyses (HR=0.56, 95% CI=0.32-0.97, P=3.7x10^{-2}) (**Table 4.4**). Patients carrying the minor (A) allele had a median increase in survival of 153 days as compared to patients homozygous for the major (G) allele.

SNP	Allele	Tumour location	N	Events	MAF	INFO	HR	95% CI	*P*
rs189655236	T	Proximal	514	413	0.0078	0.73	0.71	0.31-1.58	0.4
rs144717887	A	Proximal	514	413	0.016	0.92	0.56	0.32-0.97	**3.7x10^{-2}**
rs698022	T	Distal	493	358	0.089	0.83	0.96	0.73-1.26	0.78

Table 4.4. Replication of previously reported SNP associations with survival. Independent replication was carried out using patients from COIN and COIN-B. I had 54, 71 and 84% power to replicate the associations for rs189655236, rs144717887 and rs698022, respectively. Minor allele, tumour location, sample size, number of events, minor allele frequency (MAF) and imputation score (INFO) are shown for each

SNP as well as the Hazard ratio (HR), 95% confidence intervals (CI) and *P*-value for multivariate analyses.

4.4 Discussion

4.4.1 Independent loci replicated in the UK Biobank

I considered the relationship between inherited genetic variation and survival by location of the CRC. rs76011559 lies 123.5kb upstream of *CUL1* and replicated as a prognostic biomarker in patients with proximal tumours. *CUL1* encodes Cullin1 a member of the Cullin protein family which provides a scaffold for the ubiquitin ligase E3, mediating the degradation of proteins involved in signal transduction, transcription and cell cycle progression. As a consequence, Cullin1 regulates the cell cycle, cell proliferation, invasion, migration and metastasis (Wang et al. 2017a) and upregulation of Cullin1 in CRC tissue is a negative prognostic biomarker (Wang et al. 2015; Wang et al. 2017a; Wang et al. 2017b). However, rs76011559 was not an eQTL for *CUL1* so further studies are necessary to determine the regulatory mechanism for this SNP.

rs12273047 at 11p15.4 was also replicated in patients with rectal tumours; however, this SNP was intergenic with no clear mechanisms of action. Studies have suggested that affected genes can be up to 2Mb away from the associated SNPs and that these intergenic SNPs can often be surrounded by large insertions/deletions and act as markers of large scale genomic changes (Brodie et al. 2016). As an example, one study of the CRC predisposition SNP rs6983267 at 8q24 implicated the gene *MYC*, 335Kb downstream from rs6983267, via regulation of the transcription factor TCF4 (Tuupanen et al. 2009).

4.4.2 *PI4K2B* expression may be a prognostic biomarker for distal CRC

PI4K2B was associated with survival in patients with distal cancers beyond the threshold for multiple testing and the lead SNP rs313566 was not associated with survival in patients with proximal or rectal tumours – suggesting anatomical specificity. I sought further mechanistic understanding of this SNP. rs313566 was an eQTL for *PI4K2B* in several cell types with the A-allele associated with increased expression. Interestingly, I found that higher *PI4K2B* expression in tumour tissue was associated with improved survival in patients with colorectal tumours from THPA. *PI4K2B* encodes a member of the type II PI4 kinase protein family, responsible for overall PI4-kinase activity of the cell and PI4KII beta depletion has been associated with a more invasive phenotype in minimally invasive cell lines (Alli-Baloguna et al. 2016). However, I failed to replicate the association between rs313566 and survival in UKB patients with distal tumours, possibly due to the lack of clinicopathological factors available for inclusion in the regression models and the mixed staging of CRC patients in the UKB dataset; further studies are therefore necessary to substantiate our observations.

4.4.3 Replication of a previously reported prognostic SNP

Labadie *et al.* (2022) reported on a genome wide search for prognostic SNPs in the ISACC cohort (Chapter 2, Section 2.3.3). No loci were significantly associated with disease specific survival in the full cohort or stage-stratified analyses. However, 3 independent variants showed a significant association when stratified by location of the primary tumour. I found that rs144717887 at 14q31.3 replicated with the same direction of effect in a multivariate analysis of COIN and COIN-B and represents a

potential prognostic biomarker for proximal CRCs. However, rs144717887 sits in a low-LD intergenic region with no clear mechanism of action, so further study of potential long-range mechanisms is required.

4.4.4 Significant gene-sets

The gene-sets ‘negative regulation of phospholipid biosynthetic process’ and ‘phosphatidic acid biosynthetic process’ remained significant in our meta-analyses in patients with rectal cancers. Phospholipids have a wide range of physiological functions, including forming the cell membrane, regulating apoptosis and mitochondrial physiology, and phospholipid-derived messenger molecules are involved in intra and extra-cellular signalling. Interestingly, total amount of phospholipids in the cell membrane has been associated with cancer transformation of the cell, with differences in phospholipid composition being predictive of CRC metastases (Dobrzynska et al. 2005). Phosphatidic acid (PA) is the smallest and simplest phospholipid. PA is an important molecule for the stability and activity of the mTOR complex, a protein kinase that suppresses apoptotic signals in cancer cells (Foster 2009). These associations are intriguing given their probable biology and are candidates that warrant further investigation.

Chapter 5: Germline variation in RAS Protein Activator Like 2 may predict survival in patients with *RAS*-activated colorectal cancer

5.1 Introduction

5.1.1 Treatments for *RAS* mutant CRC

Monoclonal antibodies against EGFR, such as cetuximab, have shown benefit in *KRAS* and *RAS*, wild-type advanced CRC when either used as a monotherapy (Karapetis et al. 2008b; Guren et al. 2017) or in combination with chemotherapy (Khattak et al. 2015; Stintzing et al. 2016; Li et al. 2020) (Chapter 1, Section 1.1.3.5). In contrast, targeted treatments for patients with *RAS* mutant disease are only just emerging (Porru et al. 2018; Meng et al. 2021). Given that around half of all CRCs are *RAS* mutant, this represents a clear unmet clinical need. AMG 510 (Sotorasib), an inhibitor of *KRAS* G12C, traps mutant KRAS in its inactive GDP-bound state (Lito et al. 2016) and has shown effectiveness in a phase 2 trial of non-small cell lung cancer (Skoulidis et al. 2021). MRTX849 (Adagrasib) also binds KRAS G12C and inhibits intercellular signalling (Hallin et al. 2020), and has shown promising efficacy in patients with colorectal, non-small cell lung, endometrial, pancreatic and ovarian cancers (Sabari et al. 2021). However, both treatments are only effective in cancers harbouring G12C, which occurs in just 1-3% of CRCs. Identifying drug targets for improved survival in patients with *RAS* mutant CRC therefore remains challenging.

5.1.2 This study

Relating germline variation to outcome in patients with *RAS* mutant cancers offers the prospect of identifying novel therapeutic targets. To explore this possibility, I analysed GWAS and survival data on patients with advanced CRC from COIN and COIN-B (Chapter 2, Section 3.1). Patients' tumours were profiled for mutations in the mitogen-activated protein kinase (MAPK) and Akt pathways, to help stratify my survival analyses by MAPK pathway activation status.

5.2 Materials and Methods

5.2.1 Patients and samples

Of the 2,671 patients recruited to COIN and COIN-B, 1,948 had germline genotyping and survival data available. The minimum MAF for SNPs was set at 5% leaving 2.9 million SNPs for analysis. See Chapter 2, Section 2.3 for full details on patients, DNA extraction, genotyping and QC. See Chapter 2, Section 2.3.1.5 for details on measurements for response to treatment.

5.2.2 Somatic genotyping

Tumour samples were not available, or were of insufficient quantity, in 301 of the 1,948 patients (Chapter 2, Section 3.1.5). Overall, *KRAS* mutations were identified in 637/1589 (40.1%), *NRAS* mutations in 54/1546 (3.5%), *BRAF* mutations in 143/1554 (9.2%) and *PIK3CA* mutations in 212/1448 (14.6%) CRCs. MSI was detected in 45/1237 (3.6%) CRCs (Smith et al. 2013). Of those also tested for *BRAF* mutations, 13/45 (28.9%) CRCs with MSI carried *BRAF* V600E as compared with 93/1185 (7.8%) without MSI (P=3.1x10^{-6}), consistent with their sporadic nature (Lao and Grady 2011).

5.2.3 Patients with MAPK-activated CRC

MAPK-activated CRCs were assigned as those carrying *KRAS*, *BRAF* or *NRAS* mutations. In total, 829 patients with MAPK-activated CRCs had corresponding GWAS data. I excluded patients with potentially Akt-activated tumours (those with *PIK3CA* mutations, n=108), MSI (n=20) and those in whom covariate data was lacking (n=7 for

platelet count, primary tumour surface area, time to metastases or synchronous/ metachronous metastases). Of the remaining 694 patients, 521 (75.1%) carried *KRAS* mutations, 44 (6.3%) *NRAS* mutations, 120 (17.3%) *BRAF* mutations and 9 (1.3%) had combinations of these mutations (**Figure 5.1, Table 5.1**). For comparison, I analysed 760 patients without MAPK-activated tumours (i.e. those with *KRAS*, *NRAS* and *BRAF* wild-type CRC) and a further subset whose CRCs carried *PIK3CA* mutations as a marker of Akt-activation (n=87 patients with covariate data).

5.2.4 Statistical analyses

I previously identified clinicopathological factors associated with survival in patients from COIN and COIN-B (Chapter 3, Section 3.1). Dimensionality reduction was performed using PCA to reduce the risk of overfitting (Chapter 2, Section 4.2) the first five were selected (but only 4 were necessary to reach the 70% variance explained threshold when analysing patients with *NRAS* mutations). I carried out the GWAS for OS under an additive model. All analyses performed by MAPK gene mutation status were multivariate.

Gene and gene-set analysis was completed on the summary statistics from the association analysis to identify genes containing significant numbers of highly associated SNPs and significantly enriched gene-sets (Chapter 2, Section 2.4.5).

5.2.5 Bioinformatic analyses

See Chapter 2, Sections 2.4.3, 2.5.1 and 2.3.5 for details on GWAS analysis, LocusZoom plots and eQTL analyses, respectively.

5.2.6 The Cancer Genome Atlas (TCGA) analyses

The TCGA database (Chapter 2, Section 2.3.6) was used to find CRC patients with LOF in *RASAL2* due to the presence of somatic *RASAL2* truncating mutations or hypermethylation of the *RASAL2* locus. Data was accessed via the TCGA data portal (https://portal.gdc.cancer.gov/exploration) and the TCGA definition of LOF simple somatic mutations (SSMs) was used. Methylation array data collected using the Illumina human methylation 450 platform was downloaded from the TCGA data repository for 345 CRC samples, containing beta coefficients for methylation levels at each of 485,578 CpG islands across the genome.

To find samples with hypermethylated *RASAL2,* a mean beta coefficient was calculated for the eight CpG islands mapping to the promoter region of *RASAL2* (chr1:178,092,729-178,093,729). Due to the distribution of the mean beta coefficient being right skewed the median absolute deviation (MAD) was chosen as a suitable statistic for extracting hypermethylated samples, these were defined as those more than 2*scaling factor (1.4826; used to approximate a normal distribution)*MAD above the median beta coefficient for the population.

Truncated *RASAL2* and hypermethylated *RASAL2* samples tested for SSMs, were screened for co-occurring oncogenic *KRAS* and *NRAS* mutations (those within codons 12, 13, 59, 61, 117 or 146) (Zheng et al. 2019).

5.3 Results

5.3.1 Clinicopathological factors in patients with and without MAPK-activated CRCs

Patients with MAPK-activated CRCs were defined as those carrying *KRAS*, *NRAS* or *BRAF* mutations and that did not have Akt-activating mutations (n=108) or MSI (n=20). After QC, 694 patients had MAPK-activated CRCs (**Figure 5.1**). Patients with MAPK-activated CRCs had more right sided primary tumours, worse response at 12-weeks and poorer survival (median OS 433 days) as compared to patients without MAPK-activated CRCs (*KRAS*, *NRAS* and *BRAF* wild-type, n=760, median OS 611 days; HR=1.57, 95% CI=1.39-1.77, P=2.6x10^{-13}) (**Table 5.1**).

5.3.2 Genome-wide analysis and power considerations

Genome-wide SNP, gene and gene-set analyses were performed to identify determinants of survival in patients with MAPK-activated CRCs using the first five principal components as covariates, which explained 71.7% of the total variance for previously established prognostic factors (Chapter 2, Section 2.4.2). I had >80% power to detect a hazard ratio of 1.61 for SNPs with MAF>0.2 (Chapter 2, Section 2.4.4). No detectable genomic inflation was observed (λ=1.08). No SNPs passed the threshold for genome-wide significance.

Following LD based clumping, SNPs at eight independent loci passed the threshold for suggestive significance. The lead SNPs, summary statistics and any genes they overlap are listed in **Table 5.2**.

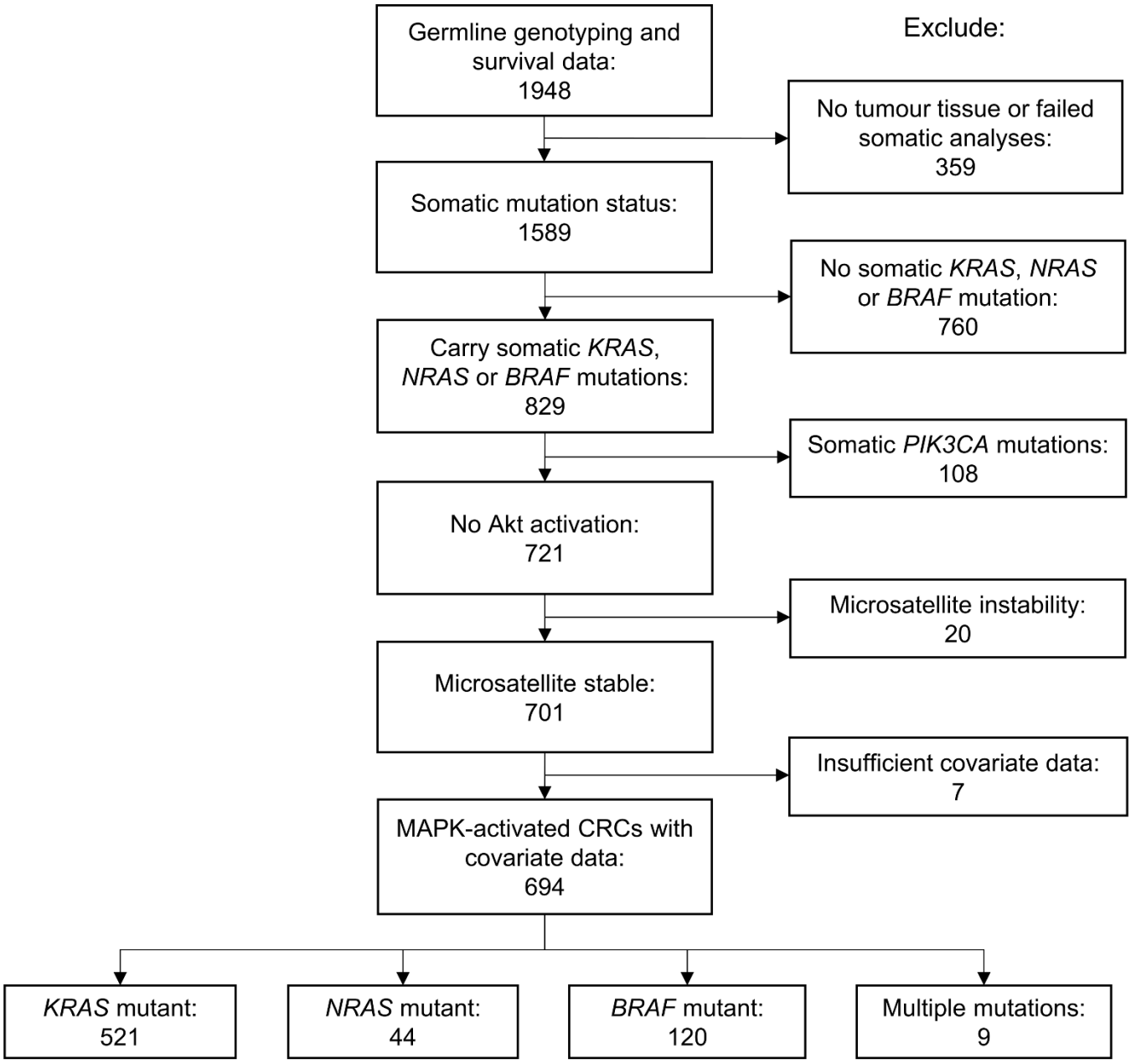


Figure 5.1. CONSORT diagram of patients with MAPK-activated colorectal cancers. Of the 1,948 patients with germline genotyping and survival data, 694 had MAPK-activated tumours without somatic *PIK3CA* mutations (no Akt activation) or microsatellite instability and had covariate data. Nine patients had CRCs with two MAPK-activating mutations (eight with *KRAS* and *NRAS* mutations and one with *KRAS* and *BRAF* mutations). 760 patients did not have MAPK-activated tumours, defined as *KRAS*, *NRAS* and *BRAF* wild-type.

Clinicopathological factor		Patients with MAPK-activated CRCs (n=694)		Patients without MAPK-activated CRCs (n=760)		*P*-value
		n	%	n	%	
Sex	Male	436	62.8	535	70.4	$2.2x10^{-3}$
	Female	258	37.2	225	29.6	
Age	Median (years)	64	-	64	-	-
Response at 12-weeks	Responders	295	50.2	452	69.0	$1.9x10^{-11}$
	Non-responders	293	49.8	203	31.0	
	No data	106		105		
Overall survival	Median (95% CI) (days)	433 (397-465)	-	611 (569-659)	-	$2.6x10^{-13}$
WHO performance status	0	330	47.6	356	46.8	$4.7x10^{-2}$
	1	301	43.4	359	47.2	
	2	63	9.1	45	6	
Site of primary tumour	Left colon	137	19.7	235	30.9	$2.1x10^{-12}$
	Right colon	233	33.6	127	16.7	
	Rectosigmoid junction	94	13.5	133	17.5	
	Rectum	219	31.6	253	33.3	
	Unknown colon	3	0.4	2	0.3	
	Multiple sites	8	1.2	10	1.3	
Status of primary tumour	Resected	400	57.6	411	54.1	0.19
	Unresected	294	42.4	349	45.9	
Surface area of primary tumour	Median (cm)	1.85	-	1.88	-	-
	Range (cm)	1.29-2.66	-	1.26-2.80	-	
Timing of metastases	Metachronous	206	29.7	241	31.7	0.44
	Synchronous	488	70.3	519	68.3	
Type of metastases	Liver only	120	17.3	199	26.2	$2.3x10^{-4}$
	Liver + others	394	56.8	386	50.8	
	Non-liver*	180	25.9	175	23	
Number of metastatic sites	1	220	31.7	290	38.2	$5.9x10^{-3}$
	2	275	39.6	301	39.6	
	≥3	199	28.7	169	22.2	
MAPK-activated		694	100	0	0	-
Mutation status	*KRAS* mutation	521	75.1	0	0	-
	NRAS mutation	44	6.3	0	0	-
	BRAF mutation	120	17.3	0	0	-
	multiple mutations	9	1.3	0	0	-

Table 5.1. Clinicopathological features of stage IV patients with and without MAPK-activated tumours. Data are n (%) or median. Differences between patients with and without MAPK-activated CRCs were analysed using a Chi-squared test, Cox regression (for overall survival) and Fisher's exact test (for stage). Response was defined as complete or partial response using RECIST 1.0 guidelines and non-response was defined as stable or progressive disease. *Non-liver metatases included those in the lungs, peritoneum and lymph nodes.

SNP	Locus	Minor allele	HR	95% CI	*P*	Genes
rs7008272	8q13.1	T	1.44	1.3-1.7	$4.7x10^{-7}$	*LINC01299*
rs78154513	6q21	T	1.50	1.3-1.8	$1.2x10^{-6}$	-
rs9592365	13q21.32	A	1.53	1.3-1.8	$1.5x10^{-6}$	-
18-56679242	18q21.31	AT	1.46	1.3-1.7	$2.6x10^{-6}$	-
rs3794586	15q14	A	0.65	0.5-0.8	$5.0x10^{-6}$	*RYR3*
rs6981227	8p23.2	G	0.69	0.6-0.8	$5.2x10^{-6}$	-
rs72623200	2q31.1	C	0.51	1.3-1.8	$5.5x10^{-6}$	*CCDC173*
rs17282574	11q21	G	1.44	1.2-1.7	$6.2x10^{-6}$	-

Table 5.2. Lead single nucleotide polymorphisms (SNPs) from independent loci that reached suggestive significance in a multivariate analysis of overall survival in patients with MAPK-activated advanced CRC (n=694). Cytogenic band, minor allele, *P*-value, hazard ratio and 95% confidence intervals are shown for overall survival. Genes overlapping with the SNPs attributed to each locus are listed. The SNP at locus 18q21.31 has yet to be assigned an rs ID and so is named by Chromosome-base pair.

5.3.3 Gene level association analysis

In MAGMA gene analysis, RAS Protein Activator Like 2 (*RASAL2*) at 1q25.2, was the most significant gene associated with survival in patients with MAPK-activated CRCs (P=2.0x10^{-5}) (**Figure 5.2**), although it did not achieve formal genome-wide significance. Patients carrying the minor (A) allele in the lead SNP, rs12028023 in intron 1 of *RASAL2*, had a median increase in survival of 167 days as compared to patients carrying the major (G) allele (HR=0.63, 95% CI=0.5-0.8, P=1.3x10^{-5}, **Figure 5.3**). In contrast, rs12028023 genotype was not associated with survival in patients without MAPK-activated tumours (HR=1.00, 95% CI=0.81-1.23, P=0.98) nor a subset whose CRCs carried *PIK3CA* mutations as a marker of Akt-activation (HR=1.72, 95% CI=0.87-3.37, P=0.12); the difference in the relationship between patient groups was significant ($P_{Z\text{-}test}$=2.1x10^{-3} and 5.3x10^{-3}, respectively). Cetuximab administration did not influence the prognostic effect of rs12028023, regardless of the MAPK-activation status (MAPK-activated $P_{Z\text{-}test}$=0.29, non-activated $P_{Z\text{-}test}$=0.49).

The rs12028023 A-allele was also associated with improved response at 12-weeks in patients with MAPK-activated cancers (77/128, 60.2% of patients carrying the A allele responded compared to 212/447, 47.4% with the G allele, OR=1.62, 95% CI=1.11-2.36, P=1.2x10^{-2}). This relationship was not seen in patients without MAPK activated cancers (93/134, 69.4% versus 352/513, 68.6%, OR=0.98, 95% CI=0.70-1.51, P=0.91).

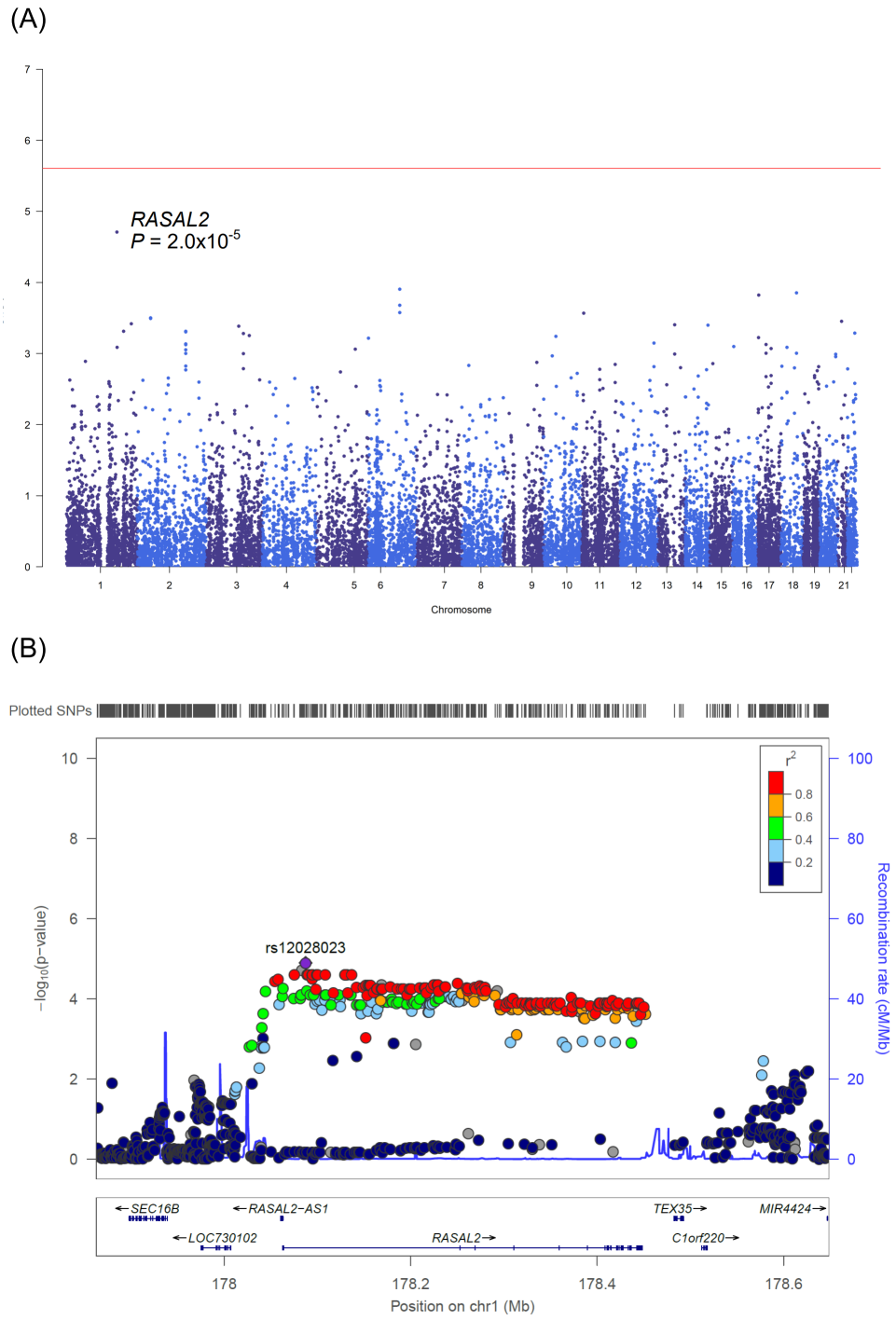


Figure 5.2. Relationship between gene, genotype and survival in 694 patients with MAPK-activated colorectal cancers. (A) Manhattan plot of gene associations with overall survival (OS). Genes are ordered by chromosome position and plotted against the -$\log_{10}(P)$ for their association with OS. The red line represents the threshold for genome-wide significance (P=2.5x10^{-6}). **(B)** Regional locus zoom plot shows results of the analysis for single nucleotide polymorphisms (SNPs) and recombination rates. $-\log_{10}(P)$ (y axis) of the SNPs are shown according to their chromosomal positions (x axis) for an area 200Kb upstream and downstream of *RASAL2*. The sentinel SNP (purple) is labelled by its rsID. The colour intensity of each symbol reflects the extent of linkage disequilibrium with the sentinel SNP, deep blue (r^2=0) through to dark red (r^2=1.0). Genetic recombination rates, estimated using 1000 Genomes Project samples, are shown with a blue line. Physical positions are based on NCBI build 37 of the human genome. Also shown are the relative positions of genes and transcripts mapping to the region of association. Genes have been redrawn to show their relative positions; therefore, maps are not to physical scale.

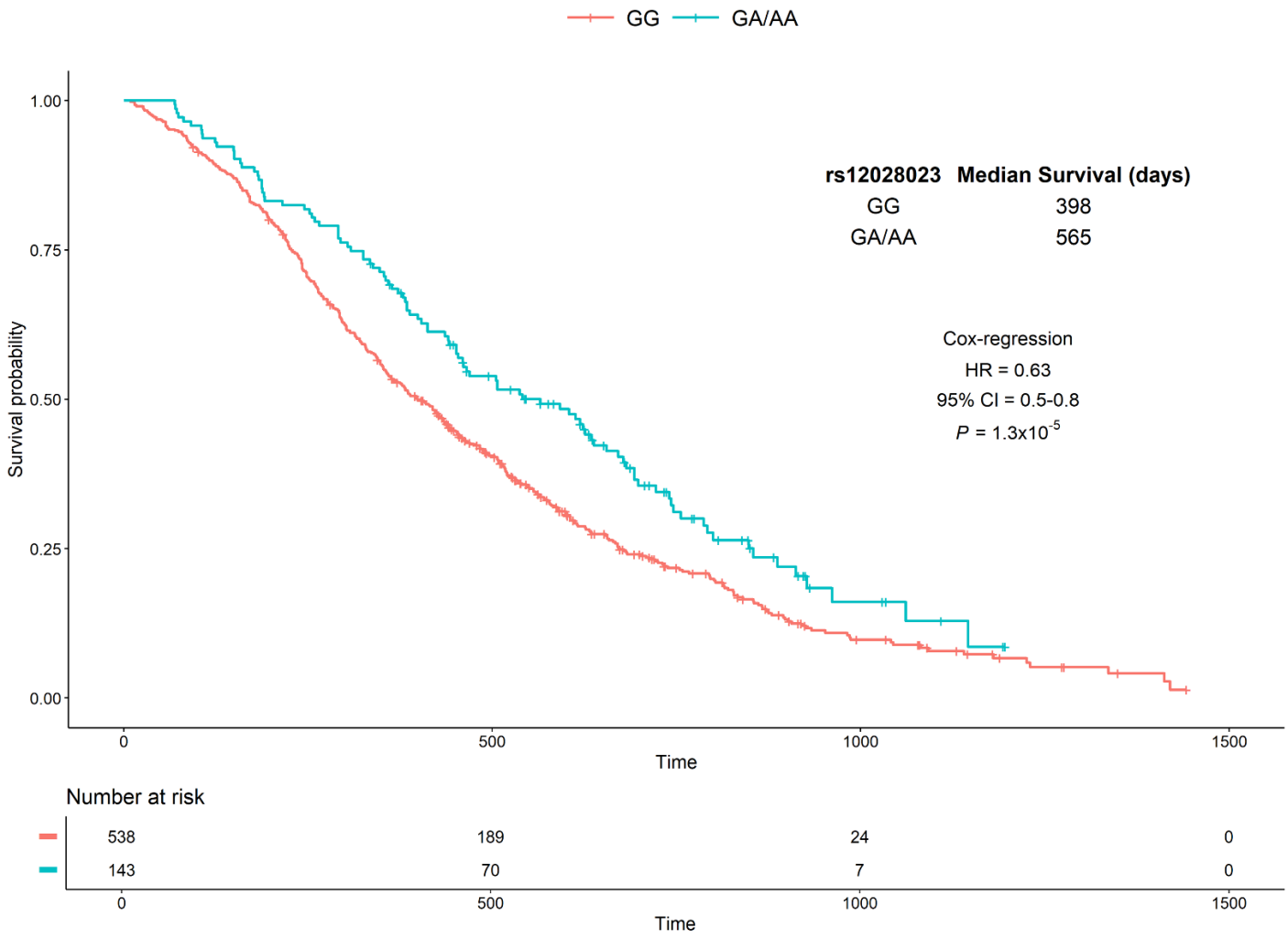


Figure 5.3. Kaplan-Meier plot of the relationship between rs12028023 genotype and overall survival in patients with MAPK-activated colorectal cancers. Time in days plotted against survival probability for patients homozygous for the major allele (GG) and heterozygous (GA) or homozygous for the minor allele (AA). The number of patients still at risk at each time point is shown beneath.

5.3.4 Analysis of *RASAL2* by MAPK gene mutation status

I dissected the prognostic role of *RASAL2* by MAPK gene mutation status. The rs12028023 A-allele was associated with improved survival in patients with *KRAS* (median increase of 191 days, HR=0.63, 95% CI=0.5-0.8, P=1.0x10^{-4}) and *NRAS* (median increase of 407 days, HR=0.22, 95% CI=0.05-0.9, P=3.8x10^{-2}) mutant CRCs (combined *RAS* mutant - median increase of 186 days, HR=0.62, 95% CI=0.5-0.8, P=3.4x10^{-5}), but not in patients with *BRAF* mutant CRCs (HR=1.05, 95% CI=0.6-1.8, P=0.87) (**Table 5.3**, **Figure 5.4**). Although there was a trend for a predictive effect on *RAS* compared to *RAF* mutant backgrounds, this did not reach statistical significance (for *KRAS versus BRAF* mutant, $P_{Z\text{-}test}$=0.097, *NRAS versus BRAF* mutant, $P_{Z\text{-}test}$=4.6x10^{-2}, combined *RAS versus BRAF* mutant, $P_{Z\text{-}test}$=0.085).

5.3.5 Analyses of rs12028023 as a biomarker of proliferation

I determined whether rs12028023 was associated with cell proliferation. The rs12028023 A-allele was associated with reduced surface area of the primary tumour (Beta=-0.037, SE=0.017, P=3.2x10^{-2}) in patients with MAPK-activated CRCs. This association was not observed in patients without MAPK-activated tumours (Beta= 0.016, SE= 0.017, P=0.36; $P_{Z\text{-test}}$=2.4x10^{-2}).

Group	N	HR	95% CI	*P*	Median increase in OS (days)
MAPK-activated	694	0.63	0.5-0.8	$1.3x10^{-5}$	167
KRAS mutant	521	0.63	0.5-0.8	$1.0x10^{-4}$	191
NRAS mutant	44	0.22	0.05-0.9	$3.8x10^{-2}$	407
BRAF mutant	120	1.05	0.6-1.8	0.87	-

Table 5.3. Association of the rs12028023-A allele with overall survival in patients with MAPK-activated CRC (n=694) and by somatic mutation status. Hazard ratio, 95% confidence intervals, *P*-value, and median increase in OS (days) are shown.

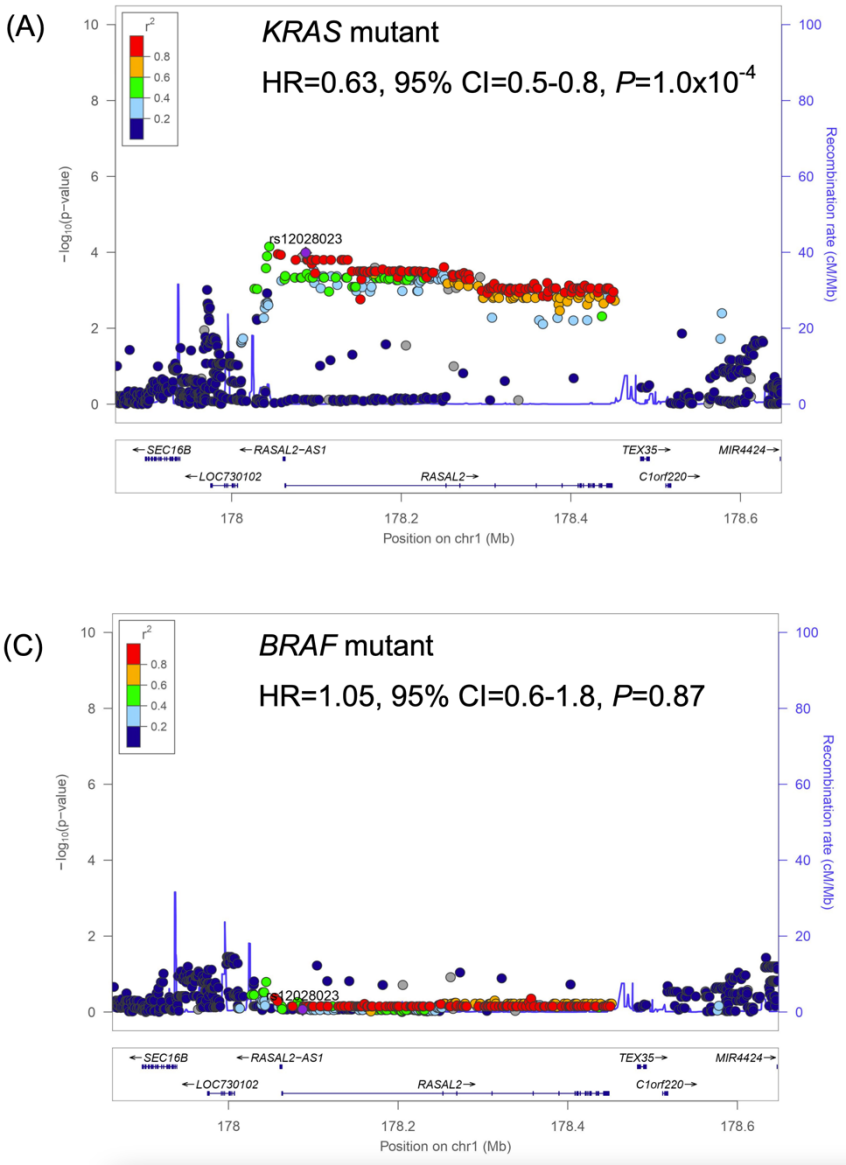


Figure 5.4. Relationship between inherited genetic variation in *RASAL2* and survival by MAPK gene mutation status. Regional Locus zoom plots for single nucleotide polymorphism (SNP) associations with overall survival in patients with colorectal cancers carrying (A) *KRAS* mutations (n=521), (B) *NRAS* mutations (n=44) and (C) *BRAF* mutations (n=120). Plots show results of the analysis for SNPs and recombination rates. $-\log_{10}(P)$ (y axis) of the SNPs are shown according to their chromosomal positions (x axis) for an area 200Kb upstream and downstream of *RASAL2*. The sentinel SNP (purple) is labelled by its rsID. The colour intensity of each symbol reflects the extent of linkage disequilibrium with the sentinel SNP, deep blue (r^2=0) through to dark red (r^2=1.0). Genetic recombination rates, estimated using 1000 Genomes Project samples, are shown with a blue line. Physical positions are based on NCBI build 37 of the human genome. Also shown are the relative positions of genes and transcripts mapping to the region of association. Genes have been redrawn to show their relative positions; therefore, maps are not to physical scale. Hazard ratio (HR), 95% confidence intervals (CI) and *P*-values are given for rs12028023.

5.3.6 Relationship between rs12028023 and *RASAL2* expression

rs12028023 was an eQTL for *RASAL2* in cultured fibroblasts from the GTEx project v8 database (P=1.6x10^{-11}) with the A-allele associated with decreased *RASAL2* expression (**Figure 5.5**). No significant association with expression was observed in the transverse (P=0.2) or sigmoid (P=1.0) colon.

5.3.7 Investigating the relationship between somatic *RASAL2* inactivation and oncogenic *RAS* mutations

I sought a (negative) correlation between *RASAL2* inactivation and *RAS* oncogenic mutations in colorectal tumours to determine whether these were mutually exclusive mechanisms for pathway activation. I considered LOF SSMs (defined by TCGA) and hypermethylation of the *RASAL2* promoter region as mechanisms of *RASAL2* inactivation.

Six hundred and sixty-nine patient CRCs from TCGA were tested for SSMs of which 33 (4.9%) had somatic *RASAL2* mutations. Of these, 6 were considered LOF (3 with a deletion resulting in the K389Rfs*7 frameshift, 1 with the G429* stop-gain mutation, 1 with the R1147* stop-gain mutation and 1 with an insertion causing the E338Gfs*70 frameshift).

To ensure that hypermethylation of *RASAL2* was exclusive to this gene and the samples were not experiencing CIMP, samples with extremely high levels of methylation across the genome were removed. A mean beta coefficient was calculated

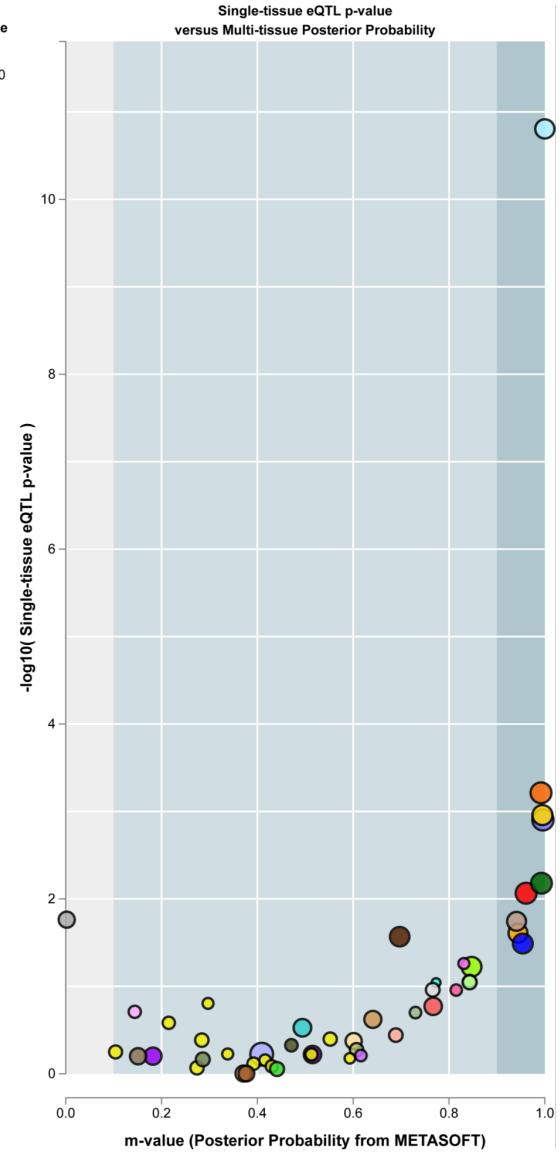
from every CpG island for each sample (n=345), approximating a normal distribution. Those samples with mean beta greater than 1 standard deviation above the mean for the population were classified as CIMP and removed from further analysis (n=41). The success of this approach was checked by comparing the co-occurrence of the *BRAF* V600E mutation, which is highly associated with CIMP (Travaglino et al. 2019). A one way two proportion Z-test showed that a significantly greater proportion of samples in the CIMP group had the *BRAF* V600E mutation than in the non-CIMP group (17/40, 42.5% versus 16/232, 0.069% respectively, P=5.1x10^{-10}).

Of the 229 CRC samples from TCGA without CIMP and that were somatically profiled, 7 had hypermethylation of CpG islands mapping to the promoter region of *RASAL2* (**Figure 5.6)**. Therefore, combined with LOF SSMs, 13 patients had CRCs with predicted inactivated *RASAL2*.

Four out of the 13 patients with inactivated *RASAL2* had co-occurring oncogenic somatic *RAS* mutations (30.8%, 2 in *KRAS* and 2 in *NRAS*). In comparison, 106/247 patients without *RASAL2* inactivation had co-occurring somatic *RAS* mutations (42.9%, 99 in *KRAS* and 8 in *NRAS*, **Table 5.4**). A one way two proportion Z-test under the alternative hypothesis of less oncogenic *RAS* mutations in the *RASAL2* inactivated group showed this to be insignificant (P=0.19).

(A)

Tissue	p-value	m-value
Brain - Spinal cord (cervical c-1)	0.2	0.297
Testis	0.02	0.00200
Brain - Hypothalamus	0.3	0.215
Ovary	0.2	0.144
Brain - Cerebellum	0.4	0.284
Brain - Amygdala	0.6	0.338
Brain - Cerebellar Hemisphere	0.8	0.392
Brain - Caudate (basal ganglia)	0.6	0.104
Esophagus - Gastroesophageal Junction	0.6	0.151
Heart - Atrial Appendage	0.6	0.182
Spleen	0.7	0.286
Adrenal Gland	0.9	0.441
Pancreas	1	0.377
Whole Blood	-	-
Colon - Sigmoid	1	0.371
Brain - Cortex	0.9	0.274
Brain - Hippocampus	0.8	0.430
Brain - Anterior cingulate cortex (BA24)	0.7	0.416
Muscle - Skeletal	0.6	0.409
Heart - Left Ventricle	0.6	0.515
Brain - Putamen (basal ganglia)	0.6	0.513
Small Intestine - Terminal Ileum	0.5	0.471
Breast - Mammary Tissue	0.3	0.494
Brain - Nucleus accumbens (basal ganglia)	0.4	0.552
Esophagus - Mucosa	0.03	0.697
Stomach	0.4	0.601
Liver	0.5	0.607
Colon - Transverse	0.2	0.641
Artery - Coronary	0.4	0.689
Artery - Aorta	0.2	0.767
Brain - Frontal Cortex (BA9)	0.09	0.843
Lung	0.06	0.847
Brain - Substantia nigra	0.7	0.593
Artery - Tibial	0.009	0.961
Cells - EBV-transformed lymphocytes	0.6	0.616
Minor Salivary Gland	0.2	0.730
Adipose - Visceral (Omentum)	0.02	0.944
Skin - Not Sun Exposed (Suprapubic)	0.03	0.954
Thyroid	0.007	0.993
Esophagus - Muscularis	0.02	0.941
Adipose - Subcutaneous	6.1e-4	0.992
Prostate	0.1	0.766
Skin - Sun Exposed (Lower leg)	1.2e-3	0.996
Vagina	0.1	0.815
Nerve - Tibial	1.1e-3	0.995
Pituitary	0.09	0.843
Uterus	0.05	0.831
Kidney - Cortex	0.09	0.773
Cells - Cultured fibroblasts	1.6e-11	1.00



(B)

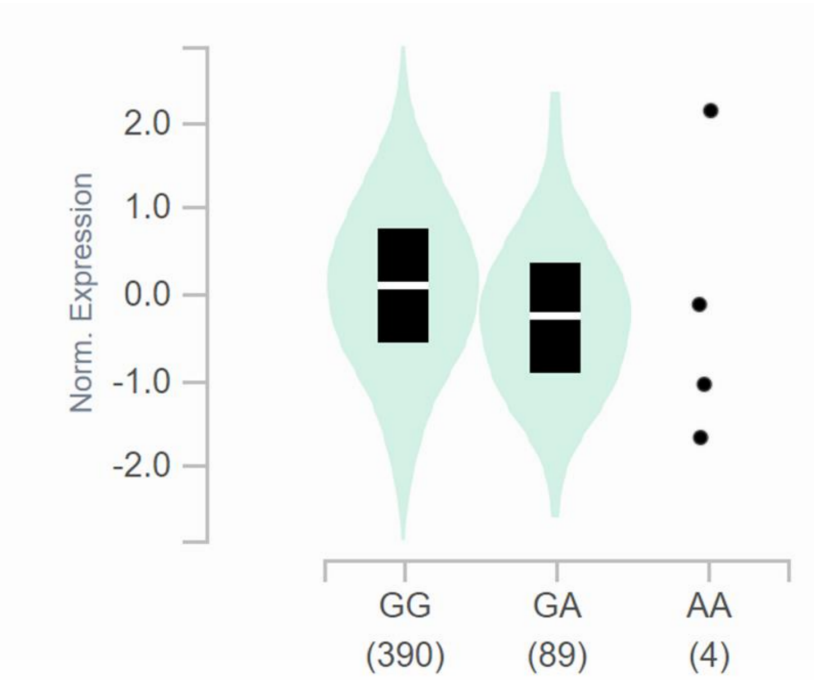


Figure 5.5. Expression quantitative trait loci (eQTL) analysis of rs12028023 for *RASAL2* expression from the GTEx database. (A) Table of *P*-values for association of the SNP and *RASAL2* expression in 49 different tissues. m-value (indicating the posterior probability that the effect is shared in each tissue tested in the cross-tissue meta-analysis, calculated by METASOFT) is plotted against -$\log_{10}$(*P*) for each tissue. **(B)** Normalised expression values for *RASAL2* by rs12028023 genotype in 483 cultured fibroblast samples.

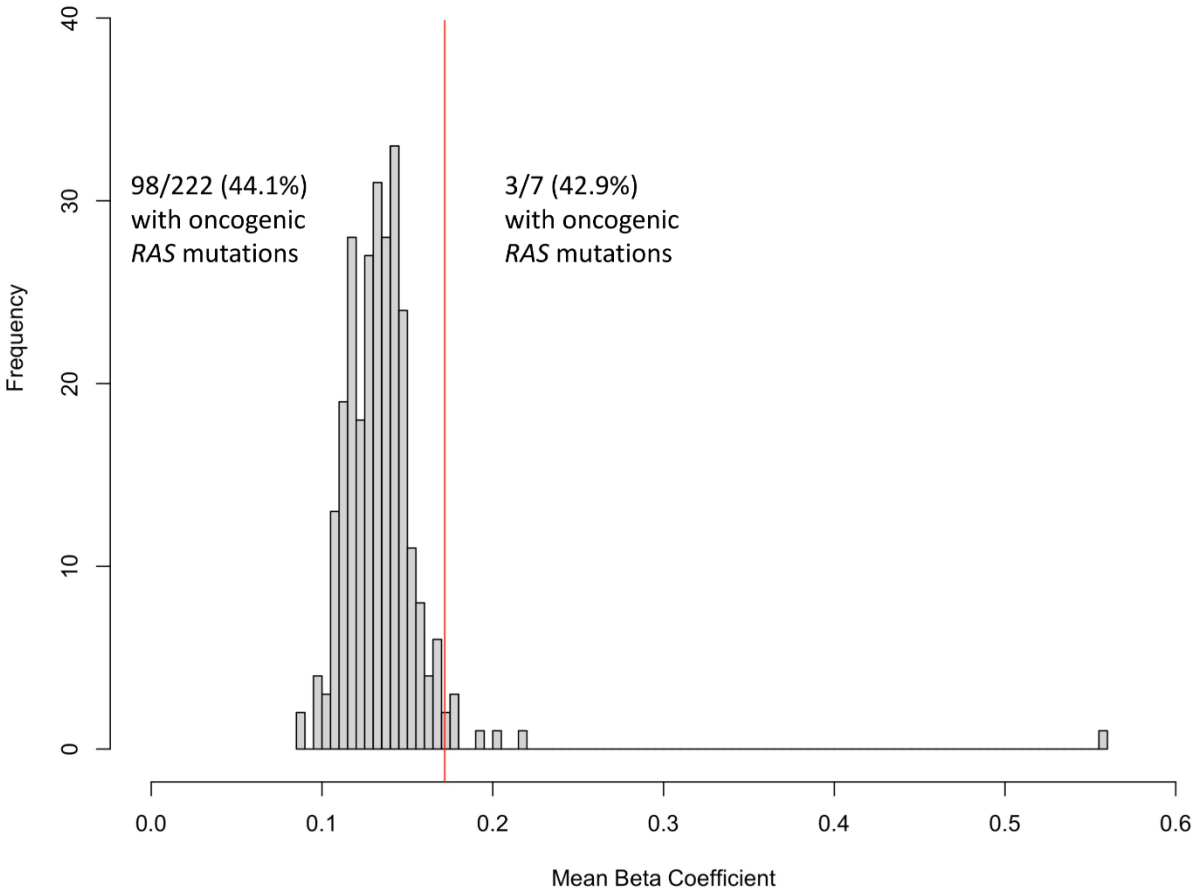


Figure 5.6. Histogram of mean methylation beta-coefficient per sample (n=304) for CpG islands mapping to the *RASAL2* promoter region. The red line is set 2*scaling factor (1.4826)*MAD above the median value, above which samples are classified as *RASAL2* hypermethylated (n=8). The frequency of the oncogenic *RAS* mutations in samples tested for simple somatic mutations from both groups are shown.

***RASAL2* status**	**n**	**Tested for SSMs n**	**Oncogenic *KRAS* mutation n**	**Oncogenic *NRAS* mutation n**	**Total oncogenic *RAS* mutations n**	**% With oncogenic RAS mutation**
LOF mutation	6	6	0	1	1	16.7%
Hypermethylated	8	7	2	1	3	42.9%
Combined	**14**	**13**	**2**	**2**	**4**	**30.8%**
Non-LOF mutation	27	27	9	0	9	33.3%
Not hypermethylated	296	222	90	8	98	44.1%
Combined	**321**	**247**	**99**	**8**	**106**	**42.9%**

Table 5.4. Co-occurrence of oncogenic *RAS* (*KRAS and NRAS*) mutations with *RASAL2* inactivation. Shows samples tested for simple somatic mutations (SSMs) with inactivated *RASAL2* (loss of function [LOF] or hypermethylated) and non-inactivated *RASAL2* from TCGA database. Combined and total groups contain only unique samples.

5.3.8 Gene-set enrichment analysis

MAGMA gene-set enrichment analysis identified five gene-sets (Golgi cisterna membrane, cisterna and stack, monoamine transport and Cul4A-RING E3 ubiquitin ligase complex) significantly associated with survival in patients with MAPK-activated CRCs after adjusting for multiple testing (q<0.05, **Table 5.5**).

GO term	Gene-set name	N genes	*P*	q
GO:0032580	Golgi cisterna membrane	23	1.0×10^{-7}	8.1×10^{-4}
GO:0031985	Golgi cisterna	49	2.0×10^{-6}	7.8×10^{-3}
GO:0015844	monoamine transport	10	1.5×10^{-5}	3.2×10^{-2}
GO:0005795	Golgi stack	68	1.6×10^{-5}	3.2×10^{-2}
GO:0031464	Cul4A-RING E3 ubiquitin ligase complex	12	2.0×10^{-5}	3.2×10^{-2}

Table 5.5. Results for MAGMA gene-set enrichment analysis. Gene-ontology (GO) term, full descriptive name, the number of genes in the gene-set, *P*-value, and false discovery rate corrected *P*-value (q) are shown. Only significant sets with q<0.05 are presented.

5.4 Discussion

5.4.1 SNPs potentially associated with survival in patients with MAPK-activated CRCs

I sought loci affecting survival in patients with MAPK-activated CRC. Of the 8 independent loci that passed the threshold for suggestive significance, 3 had overlapping genes. Of these, only Ryanodine Receptor 3 (*RYR3*) has shown previous associations with cancer. *RYR3* encodes a large protein that forms a calcium channel. rs1044129, which is not in LD with the sentinel SNP from this analysis (D'=0.0037 and R^2=0.0 in the 1000 Genomes Project European population), is in the 3'-UTR of *RYR3* and is a binding site for microRNA-367. In both breast cancer and hepatocellular carcinoma, the G allele of rs1044129 is significantly associated with increased risk and poorer overall survival (Zhang et al. 2011; Peng et al. 2015). Neither rs1044129 nor any SNP in suitable LD were included in this analysis. The mechanism of action for the sentinel SNP intronic to *RYR3* from this analysis is still unknown but warrants further study.

5.4.2 Variation in *RASAL2* may predict survival in MAPK-activated CRC

RASAL2 was the most significant gene associated with survival in patients with MAPK-activated CRCs. *RASAL2* encodes a RAS GTPase-activating protein (GAP), which negatively regulates the RAS signalling pathway by converting RAS-GTP to RAS-GDP (Pan et al. 2018). Although *RASAL2* did not pass formal genome-wide significance in our screen, its direct interaction with RAS (as one of only fourteen known RAS GAPs) (Bernards 2003) makes it an interesting candidate gene. Given that I only had 694

patients with MAPK-activated CRCs, it is more likely that I had too few cases to achieve the stringent threshold for genome-wide significance. It is noteworthy that the rs12028023 A-allele specifically improved survival in patients with *KRAS* and *NRAS* mutant cancers, but not in those with *BRAF* mutant cancers, supporting a direct effect on the upstream RAS signalling pathway. The lack of association in patients with *BRAF* mutant cancers was unlikely to be due to the small numbers of samples (n=120) since I observed this effect in a much smaller group with *NRAS* mutant cancers (n=44). Furthermore, rs12028023 did not influence survival in patients without MAPK activated CRCs, nor the subset with Akt-activation, highlighting its specificity to this pathway.

5.4.3 *RASAL2* has varying roles in colorectal cancer

In CRC, *RASAL2* inactivation promotes progression and metastasis (Jia et al. 2017) possibly via negative modulation of the *RAS* activation pathway. Zhang et al. (2019) proposed that this was due to an association with the karyopherin nuclear transport receptor family member IPO5. They showed that IPO5 is overexpressed in CRC tissue, positively associated with clinicopathological characteristics of the disease and binds to the nuclear localization sequence of RASAL2, mediating its nuclear translocation and thus removing it from the cytoplasm where it negatively regulates *RAS* pathway activation.

However, *RASAL2* has also been found to be upregulated in metastatic CRCs with higher expression associated with lymph node involvement, distant metastasis, and poorer prognosis, possibly via its involvement in the Hippo signalling pathway. *RASAL2* inhibits the expression of large tumour suppressor kinase 2, increasing the

expression of yes-associated protein 1 which is translocated to the nucleus and leads to expression of pro-proliferation genes (Pan et al. 2018).

My data suggests that *RASAL2*'s role in CRC tumorigenesis is likely to be influenced by the MAPK-activation status of the patient's cancer which was not analysed in these aforementioned studies and may help explain some of the conflicting data (Zhou et al. 2019).

5.4.4 The varying roles of *RASAL2* in other cancers

RASAL2 was identified as a tumour suppressor in prostate cancer (Min et al. 2010) where it is differentially hypermethylated, reducing expression and leading to increased cell proliferation and invasion (Tailor et al. 2021). *RASAL2* inactivation also promotes progression and metastasis in lung (Li and Li 2014) and ovarian (Huang et al. 2014) cancers via ERK regulation. In luminal B breast cancers RASAL2 loss increases MEK/ERK (extracellular regulated protein kinases) and PI3K/AKT signalling to promote invasion, as well as activating NF-kB leading to increased epithelial–mesenchymal transition (EMT) (McLaughlin et al. 2013).

However, *RASAL2* has also shown pro-oncogenic roles in triple-negative breast where its downregulation by miR-136 and miR-203 leads to suppression of cell migration, EMT and invasion (Feng et al. 2014). In hepatocellular carcinoma (HCC) *RASAL2* is hypomethylated, upregulating it and promoting invasiveness; downregulation impairs the Akt, RAS-RAF-MEK-ERK and WNT/β-catenin pathways by altering the phosphorylation of their effectors (Stefanska et al. 2014). *RASAL2* is also the target of

miR-203 in HCC, overexpression of which exhibited similar effects to *RASAL2* knockdown (Fang et al. 2017). The varying molecular pathways of *RASAL2* action in different cancers is summarised in **Figure 5.7**.

5.4.5 *RASAL2* inactivation is not correlated with somatic *RAS* mutation status

Due to *RASAL2*'s negative modulation of the MAPK pathway I hypothesized that *RASAL2* inactivation would negate the requirement for activating RAS mutations. Due to the previously reported differential methylation of *RASAL2* in HCC (Stefanska et al. 2014) and prostate cancer (Tailor et al. 2021) both hypermethylation and somatic truncating mutations were used as markers of inactivation. However, there was no significant difference in the frequency of oncogenic *RAS* mutations in CRCs from patients with or without *RASAL2* inactivation, suggesting no link between *RASAL2* inactivation and *RAS* mutation status. Therefore, polymorphisms affecting *RASAL2* expression may only have a protective effect in the presence of activating *RAS* mutations that cause aberrant regulation of the MAPK pathway.

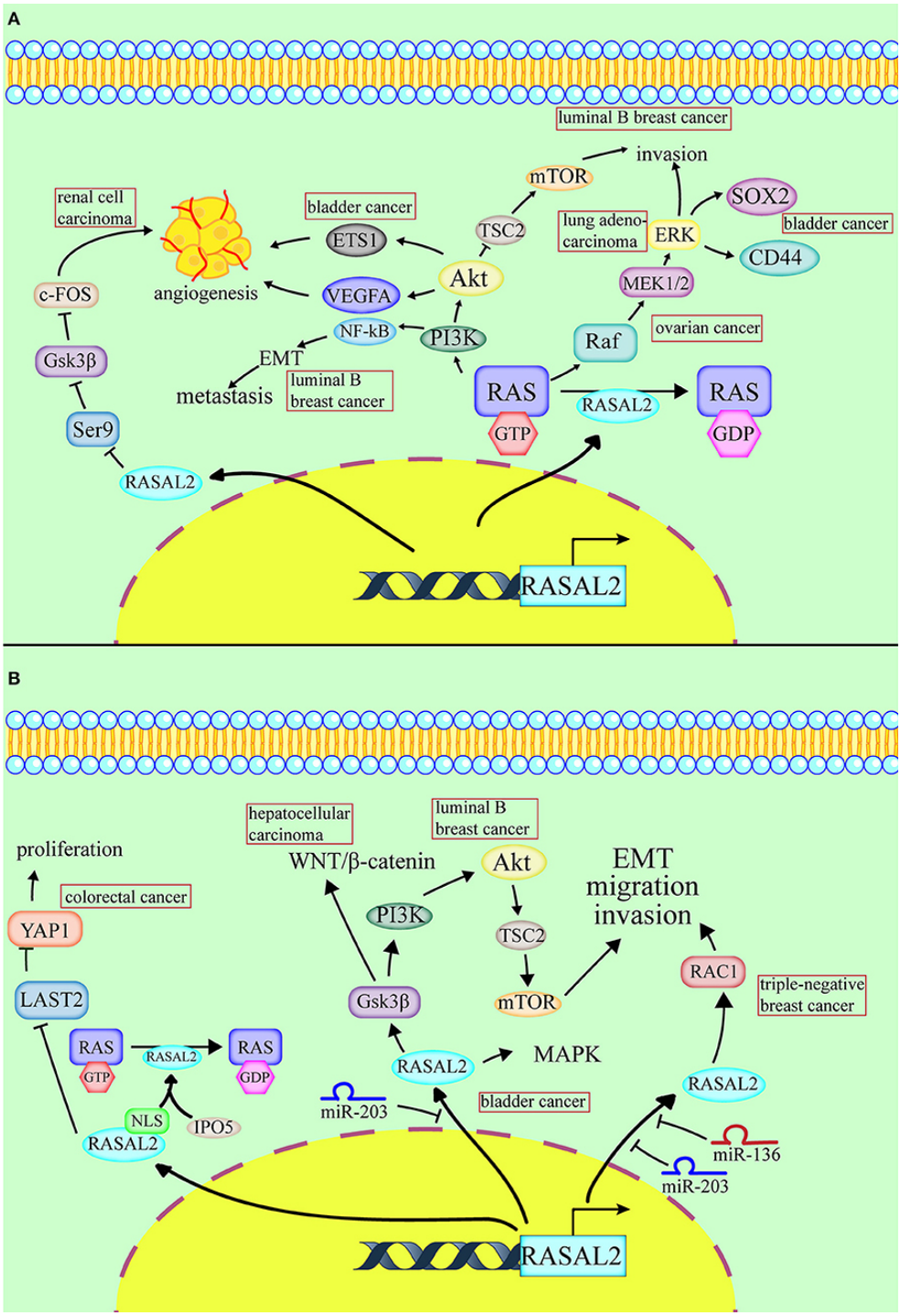


Figure 5.7. Biological roles of *RASAL2* in different cancers. (A) Renal cell carcinoma, luminal B breast cancer, bladder cancer, lung adenocarcinoma, ovarian cancer and **(B)** colorectal cancer, hepatocellular cancer, bladder cancer, Luminal B breast cancer and triple negative breast cancer. Reproduced from Zhou et al. (2019) with permission.

5.4.6 Role of differential *RASAL2* expression

Carriers of the rs12028023 A-allele were predicted to have reduced *RASAL2* expression in cultured fibroblasts, but not colonic tissue. A median increase in survival of 167 days was observed in patients with MAPK-activated CRCs and 186 days in the subset with *RAS*-mutant CRCs. Importantly, others have shown that reduced *RASAL2* expression is also associated with improved survival in two independent cohorts of patients with CRC (Pan et al. 2018), although these were not molecularly stratified by MAPK-activation status. However, these data suggest that *RASAL2* may represent a potential therapeutic target via modulation of its expression and warrant further investigation. Furthermore, given RASAL2's role in tumourigenesis in other cell types (Stefanska et al. 2014), I speculate that it may represent a target for intervention in a broader range of cancers.

5.4.7 Relationship between rs12028023 and cell proliferation

Previous research has shown knockdown of *RASAL2* in multiple CRC cell lines decreases cell proliferation, anchorage-dependent and -independent growth, cell invasion and migration (Pan et al. 2018). Interestingly, I noted that the rs12028023 A-allele was associated with reduced surface area of the primary tumour in patients with MAPK-activated CRCs, potentially supporting a link between reduced *RASAL2* expression and decreased proliferation.

5.4.8 Gene-set analysis

Five gene sets were significant for an association with OS. However, no clear link can be seen between these biological pathways and MAPK-activation. Three of the gene-sets regulate the Golgi apparatus which plays a vital role in normal cell physiology by facilitating proliferation, cell survival, migration, cellular homeostasis and cell-cell communication, all dysregulated in human cancers (Bui et al. 2021). The Cul4A-RING E3 ubiquitin ligase complex is a multi-subunit protein complex which plays a role in DNA damage repair, chromatin remodelling, DNA replication, regulation of the cell cycle, haematopoiesis, spermatogenesis, and meiosis. The sets constituent genes have shown previous associations with CRC, promoting processes like cancer progression, proliferation, and metastasis (Ren et al. 2016; Sui et al. 2017) and therefore warrant further investigation.

Chapter 6: Poly(ADP-Ribose) Polymerase Family Member 11 may predict survival in patients with wild-type colorectal cancer

6.1 Introduction

6.1.1 Somatic mutations and prognosis

Many somatic mutations in CRCs have large prognostic effects (Chapter 1, Section 1.1.4.2). *KRAS* mutations occur in approximately 40% of CRCs (Chapter 5, Section 5.1.1) and confer a significantly worse median OS (Andreyev et al. 1998; Richman et al. 2009; Eklof et al. 2013; Cremolini et al. 2015b). Mutations in *NRAS*, another member of the MAPK pathway, have also been shown to reduce median OS (Schirripa et al. 2015) but this association has not been widely replicated (Ogura et al. 2014). *BRAF* mutations are strongly associated with poorer prognosis (Tran et al. 2011a; Kalady et al. 2012), especially the V600E mutation (Guan et al. 2020). *PIK3CA* mutations are predictive of worse disease-specific survival (Kato et al. 2007), progression free survival and OS (Li et al. 2017). MSI has previously been shown to confer poor prognosis in mCRC patients (Tran et al. 2011a; Smith et al. 2013) but superior prognosis in locally advanced disease patients (Lochhead et al. 2013).

6.1.2 This study

In Chapter 3 I performed a genome wide analysis of SNP associations with OS using the COIN and COIN-B cohorts. Although I found SNPs at 17 loci suggestive of association, I considered whether the somatic genetic background was confounding our analyses and masking genome-wide significant variants. I therefore performed a

GWAS in patients with CRCs that did not have known somatic mutations affecting prognosis, together with a TWAS to support my findings.

6.2 Materials and Methods

6.2.1 Patients and samples

1,948 patients from COIN and COIN-B had germline genotyping and survival data available. The minimum MAF for SNPs was set at 5% leaving 2.8 million SNPs for analysis. See Chapter 2, Section 2.3 for full details on patients, DNA extraction, genotyping and QC.

6.2.2 Subset of patients with wild-type CRC

In Chapter 5, I identified 760 patients without MAPK-activated CRCs (those that were wild type for *KRAS*, *NRAS* and *BRAF*). Here, I further excluded patients with CRCs harbouring *PIK3CA* mutations (n=75), MSI (n=19) or that lacked somatic genetic data (n=85), leaving 581 patients (393 events) for analyses (an 'all wild-type' cohort).

6.2.3 Statistical analyses

I previously identified clinicopathological factors associated with survival in patients from COIN and COIN-B (Chapter 3, Section 3.3.1). Dimensionality reduction was performed using PCA to reduce the risk of overfitting (Chapter 2, Section 2.4.2) and the first five principal components were selected. I carried out the GWAS for OS under an additive model. Gene and gene-set analysis were completed as previously described (Chapter 2, Section 2.4.5).

Multivariate transcriptome-wide association analysis was completed using GReX imputed using whole-blood tissue MASHR-based models (Chapter 2, Section 4.6).

6.2.4 Bioinformatic analyses

See Chapter 2, Sections 2.4.3, 2.5.1 and 2.3.5 for details on GWAS analysis, LocusZoom plots and eQTL analyses, respectively.

I sought an association between Poly(ADP-Ribose) Polymerase Family Member 11 (*PARP11*) expression levels in colorectal tumours and survival in 597 CRC patients from THPA (Chapter 2, Section 2.3.7). Samples were classified as high expression using a threshold of FPKM>1.10 as per THPA recommendations (FPKM>1.64 for stage IV patient subset).

6.3 Results

6.3.1 Genome-wide analysis and power considerations

Genome-wide SNP, gene and gene-set analyses were performed to identify determinants of survival using the first five principal components as covariates, which explained 72.8% of the total variance for previously established prognostic factors (Chapter 2, Section 2.4.2). I had >80% power to detect a hazard ratio of 1.74 for SNPs with MAF>0.2 (Chapter 2, Section 2.4.4). No detectable genomic inflation was observed (λ=1.07; **Figure 6.1**).

A single SNP, rs11062901 at 12p13.32 was genome wide significant for survival in patients with all wild-type CRCs (HR=1.99, 95% CI=1.6-2.5, P=4.5x10^{-8}). Another independent SNP, rs11254422 at 10p14 was just under this threshold (HR=1.99, 95% CI=1.5-2.6, P=5.6x10^{-8}; **Figure 6.2**). Following LD based clumping, a further six independent loci passed the threshold for suggestive significance. The lead SNPs, summary statistics and any genes they overlap are listed in **Table 6.1**.

rs11062901 lies approximately 80Kb upstream of *PARP11* and carriers of the T allele had a median reduction in survival of 249 days compared to patients homozygous for the major (C) allele (**Figure 6.3**). rs11254422 lies approximately 63Kb downstream of Long Intergenic Non-Protein Coding RNA 706 (*LINC00706*) and 69Kb upstream of Long Intergenic Non-Protein Coding RNA 707 (*LINC00707*). Carriers of the A allele had a median reduction in survival of 230 days compared to patients homozygous for the major (G) allele (**Figure 6.3**).

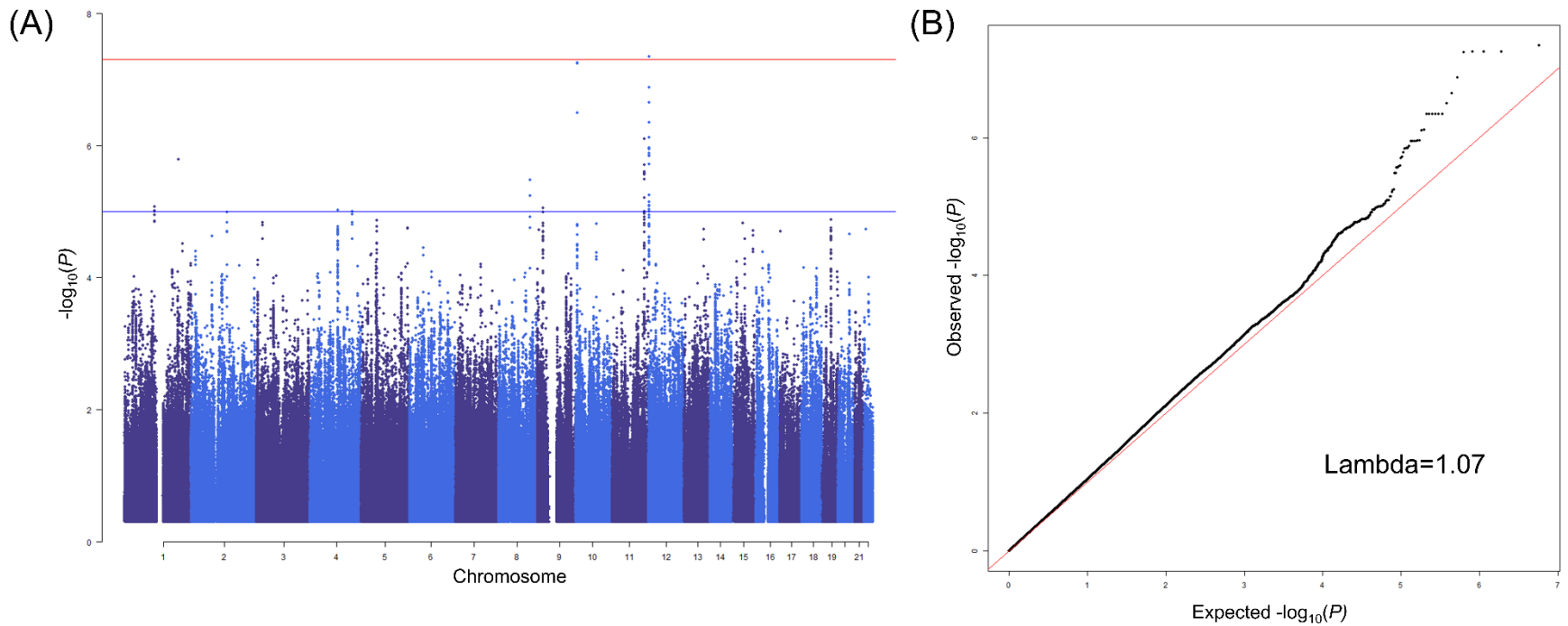


Figure 6.1. Single nucleotide polymorphism (SNP) associations with overall survival (OS) (n=581 patients with all wild-type colorectal cancer). (A) Manhattan plot: SNPs are ordered by chromosome position and plotted against the -$\log_{10}$(*P*) for their association with OS. The red line represents the threshold for genome wide significance (P=5.0x10^{-8}) and the blue line is the threshold for suggestive significance (P=1.0x10^{-5}). Covariates included the first 5 principal components representing: World Health Organisation performance status, resection status of the primary tumour, white blood cell count, platelet count, alkaline phosphatase levels, number of metastatic sites, metastases within or outside of the liver, site of primary tumour, surface area of primary tumour, time from diagnosis to metastases and metachronous versus synchronous metastases. **(B)** Quantile-quantile plot: expected -$\log_{10}$(*P*-value), under the null hypothesis of no association between genotype and OS, plotted against observed -$\log_{10}$(*P*-value).

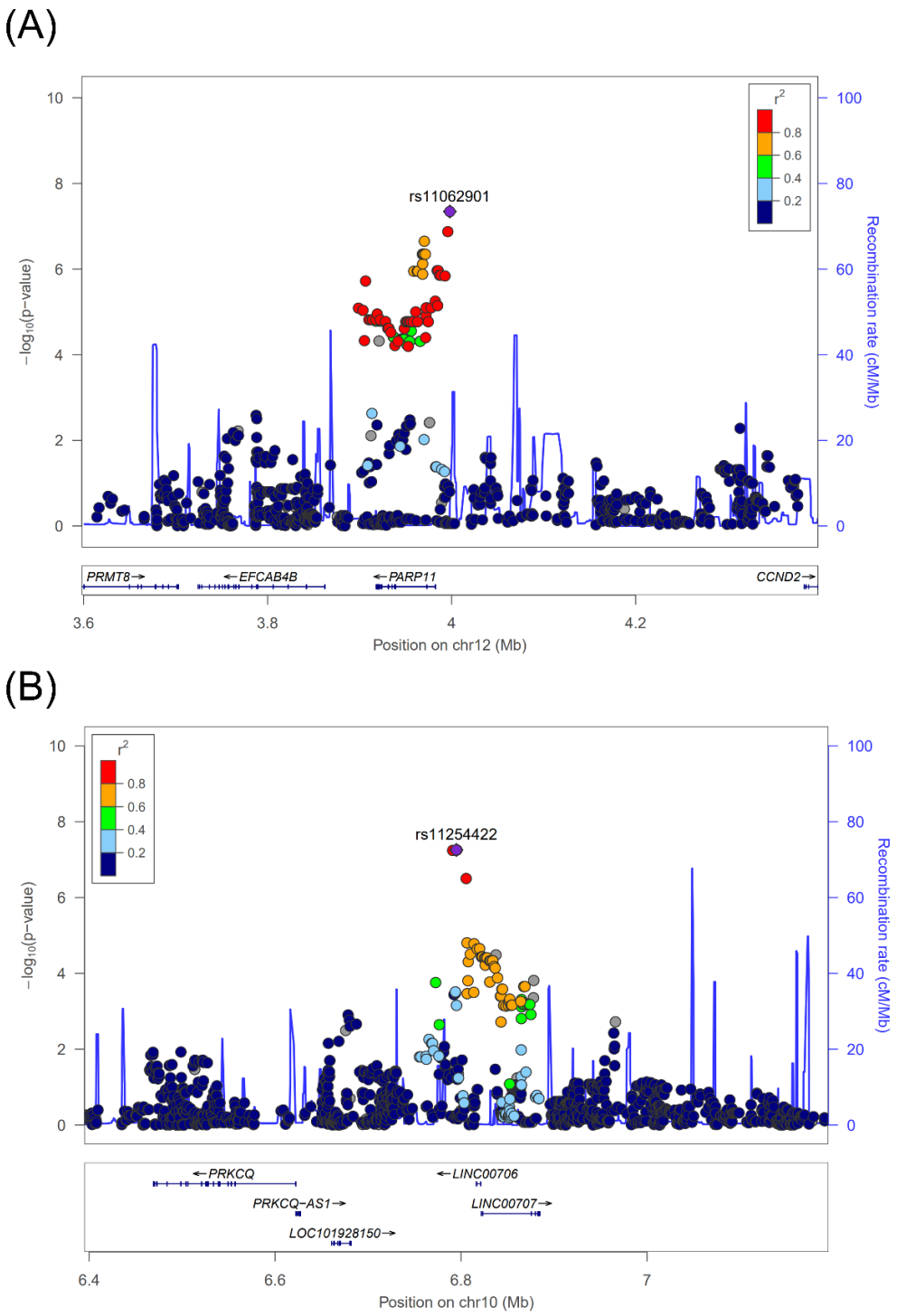


Figure 6.2. Regional locuszoom plots for the association of single nucleotide polymorphisms (SNPs) at (A) 12p13.32 and (B) 10p14 with overall survival (OS) in wild-type colorectal cancers (n=581). -$\log_{10}(P)$ (y axis) of the SNPs are shown according to their chromosomal positions (x axis) for an area 400Kb upstream and downstream of the sentinel SNPs (purple), labelled by rsID. The colour intensity of each symbol reflects the extent of linkage disequilibrium with the sentinel SNP, deep blue (r^2=0) through to dark red (r^2=1.0). Genetic recombination rates, estimated using 1000 Genomes Project samples, are shown with a blue line. Physical positions are based on NCBI build 37 of the human genome. Also shown are the relative positions of genes and transcripts mapping to the region of association. Genes have been redrawn to show their relative positions; therefore, maps are not to physical scale.

SNP	Locus	Minor Allele	MAF	HR	95% CI	*P*	Genes
rs11062901	12p13.32	T	0.060	1.99	1.6-2.5	**$4.5x10^{-8}$**	***PARP11***
rs11254422	10p14	A	0.071	1.99	1.6-2.6	$5.6x10^{-8}$	*LINC00706, LINC00707*
rs35968527	11q23.3	T	0.23	1.49	1.3-1.8	$7.9x10^{-7}$	*TECTA*
rs2820289	1q32.1	T	0.080	1.84	1.4-2.4	$1.6x10^{-6}$	*IPO9-AS1, NAV1*
rs6980997	8q23.3	G	0.16	1.58	1.3-1.9	$3.3x10^{-6}$	
rs12724483	1p13.2	G	0.28	0.69	0.6-0.8	$8.4x10^{-6}$	
rs10651937	9p21.3	G	0.27	1.43	1.2-1.6	$9.0x10^{-6}$	*FOCAD*
rs6813563	4q24	A	0.36	1.99	1.6-2.5	$9.6x10^{-6}$	*BDH2, CENPE, SLC9B1, SLC9B2*

Table 6.1 Lead single nucleotide polymorphisms (SNPs) from independent loci that reached suggestive significance in a multivariate analysis of overall survival in patients with all wild-type advanced CRC (n=581). Cytogenic band, minor allele, minor allele frequency in COIN/COIN-B, *P*-value, hazard ratio and 95% confidence intervals are shown for overall survival. Genes overlapping with the SNPs attributed to each locus are listed. rs11062901 at 12p13.32 reached the threshold for genome-wide significance ($P<5.0x10^{-8}$, in bold).

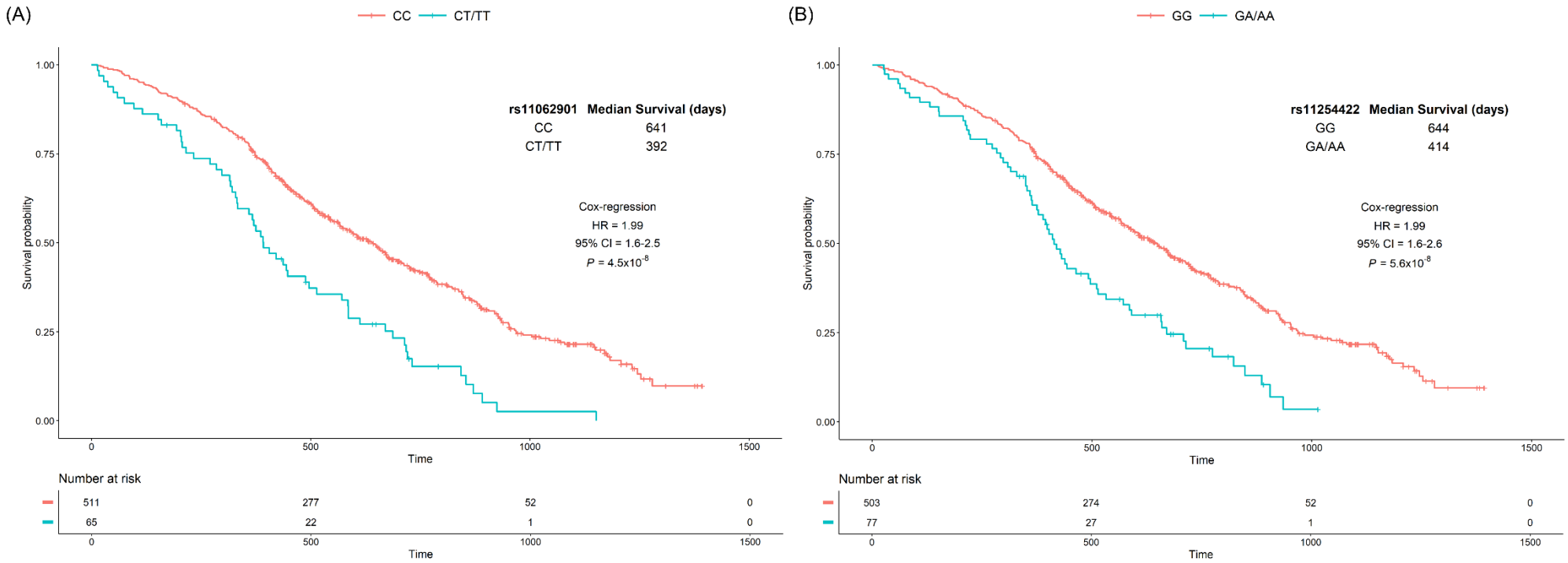


Figure 6.3. Kaplan-Meier plots for the relationship between (A) rs11062901 and (B) rs11254422 genotypes with overall survival. Time in days plotted against survival probability for patients homozygous for the major alleles and heterozygous or homozygous for the minor alleles. The number of patients still at risk at each time point is shown beneath.

6.3.2 Gene level association analysis

In MAGMA gene analysis, *PARP11* at 12p13.32, was significantly associated with OS in patients with all wild-type CRCs (P=1.4x10^{-6}; **Figure 6.4**). No gene sets were significantly associated with survival.

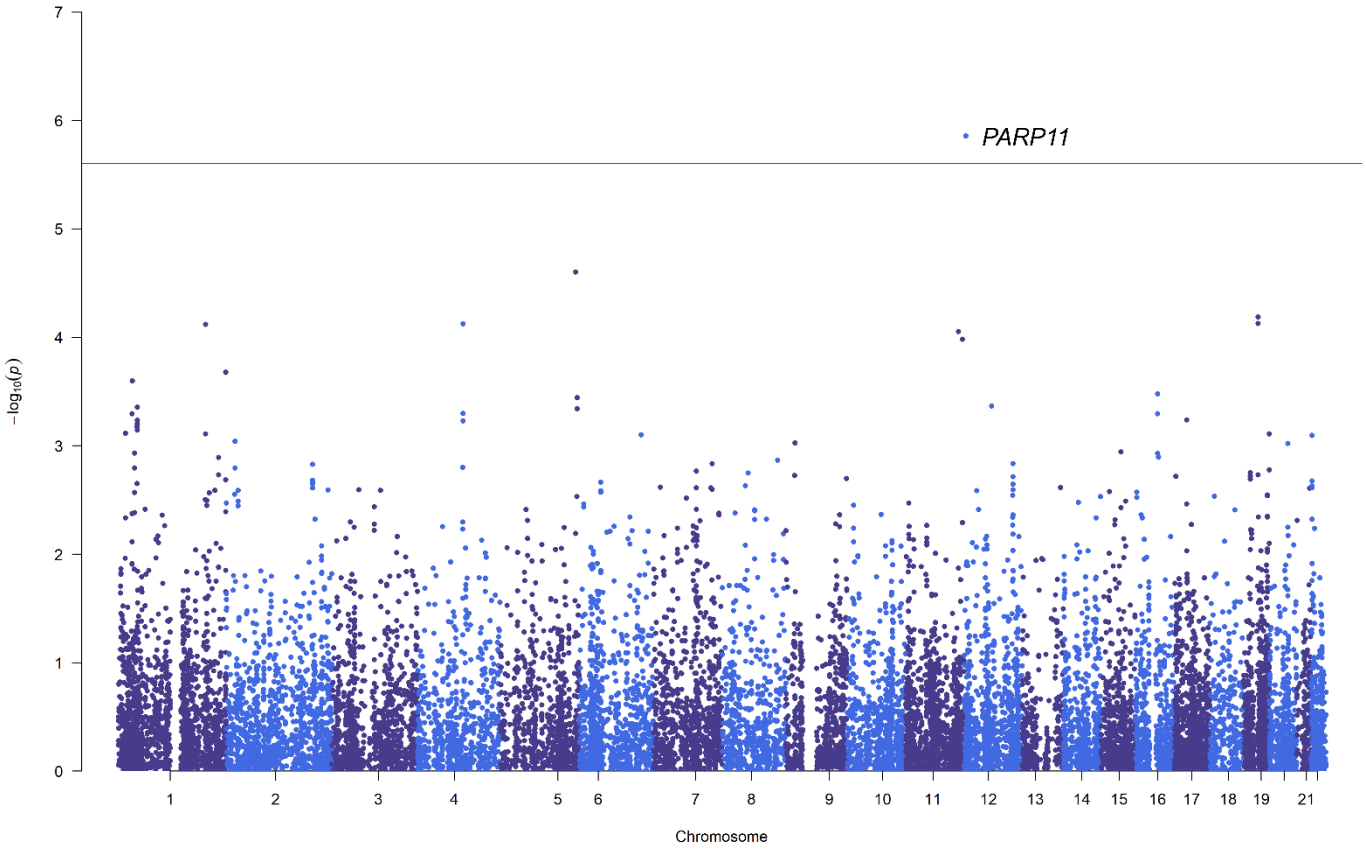
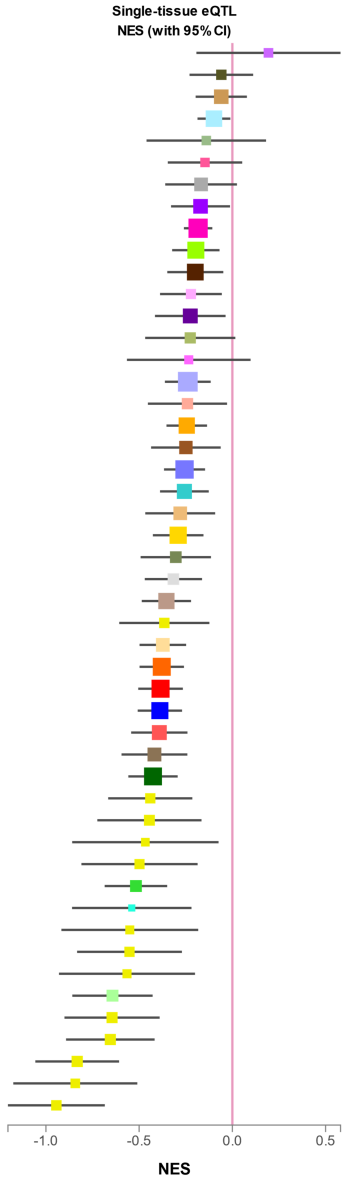


Figure 6.4. Manhattan plot of gene associations with overall survival (OS) in 581 patients with wild-type colorectal cancer. Genes are ordered by chromosome position and plotted against the -log$_{10}$(P) for their association with OS. The red line represents the threshold for genome-wide significance (P=2.5x10^{-6}).

6.3.3 eQTL analysis

rs11062901 was an eQTL for *PARP11* in 19 of the 49 tissue types tested by GTeX (based on an FDR corrected significance threshold for the specific SNP/gene combination), with the T allele being predictive of lower normalised expression (**Figure 6.5**). However, no significant association was observed in the transverse (*P*=0.40) or sigmoid ($P=3.8\times10^{-3}$) colon. rs11254422 was not an eQTL for any genes.

Tissue	Samples	NES	p-value	m-value
Cells - EBV-transformed lymphocytes	147	0.194	0.3	0.354
Small Intestine - Terminal Ileum	174	-0.0591	0.5	0.0530
Colon - Transverse	368	-0.0595	0.4	0.0430
Cells - Cultured fibroblasts	483	-0.0987	0.03	0.00
Minor Salivary Gland	144	-0.140	0.4	0.902
Vagina	141	-0.147	0.2	0.921
Testis	322	-0.168	0.09	0.948
Heart - Atrial Appendage	372	-0.171	0.04	0.930
Whole Blood	670	-0.184	2.9e-6	1.00
Lung	515	-0.196	2.7e-3	0.999
Esophagus - Mucosa	497	-0.199	0.01	0.987
Ovary	167	-0.222	0.01	0.999
Heart - Left Ventricle	386	-0.226	0.02	0.999
Liver	208	-0.226	0.07	0.973
Uterus	129	-0.234	0.2	0.905
Muscle - Skeletal	706	-0.239	1.5e-4	1.00
Artery - Coronary	213	-0.241	0.03	1.00
Adipose - Visceral (Omentum)	469	-0.245	1.4e-5	1.00
Pancreas	305	-0.249	0.009	0.989
Skin - Sun Exposed (Lower leg)	605	-0.256	6.1e-6	1.00
Breast - Mammary Tissue	396	-0.257	1.4e-4	1.00
Colon - Sigmoid	318	-0.280	3.8e-3	1.00
Nerve - Tibial	532	-0.291	3.2e-5	1.00
Spleen	227	-0.303	1.9e-3	1.00
Prostate	221	-0.316	8.0e-5	1.00
Esophagus - Muscularis	465	-0.354	2.5e-7	1.00
Brain - Frontal Cortex (BA9)	175	-0.365	3.7e-3	1.00
Stomach	324	-0.373	1.4e-8	1.00
Adipose - Subcutaneous	581	-0.379	9.7e-10	1.00
Artery - Tibial	584	-0.385	5.9e-10	1.00
Skin - Not Sun Exposed (Suprapubic)	517	-0.389	3.9e-10	1.00
Artery - Aorta	387	-0.392	6.4e-7	1.00
Esophagus - Gastroesophageal Junction	330	-0.418	5.4e-6	1.00
Thyroid	574	-0.426	6.3e-10	1.00
Brain - Hypothalamus	170	-0.440	2.0e-4	1.00
Brain - Caudate (basal ganglia)	194	-0.445	2.2e-3	1.00
Brain - Substantia nigra	114	-0.467	0.02	1.00
Brain - Hippocampus	165	-0.498	2.2e-3	0.998
Adrenal Gland	233	-0.517	7.6e-9	1.00
Kidney - Cortex	73	-0.540	1.8e-3	1.00
Brain - Spinal cord (cervical c-1)	126	-0.550	4.1e-3	0.999
Brain - Putamen (basal ganglia)	170	-0.552	1.9e-4	1.00
Brain - Amygdala	129	-0.565	3.0e-3	1.00
Pituitary	237	-0.643	2.0e-8	1.00
Brain - Nucleus accumbens (basal ganglia)	202	-0.645	1.8e-6	1.00
Brain - Cortex	205	-0.654	2.3e-7	1.00
Brain - Cerebellum	209	-0.832	1.4e-11	1.00
Brain - Anterior cingulate cortex (BA24)	147	-0.842	2.3e-6	1.00
Brain - Cerebellar Hemisphere	175	-0.944	5.8e-11	1.00



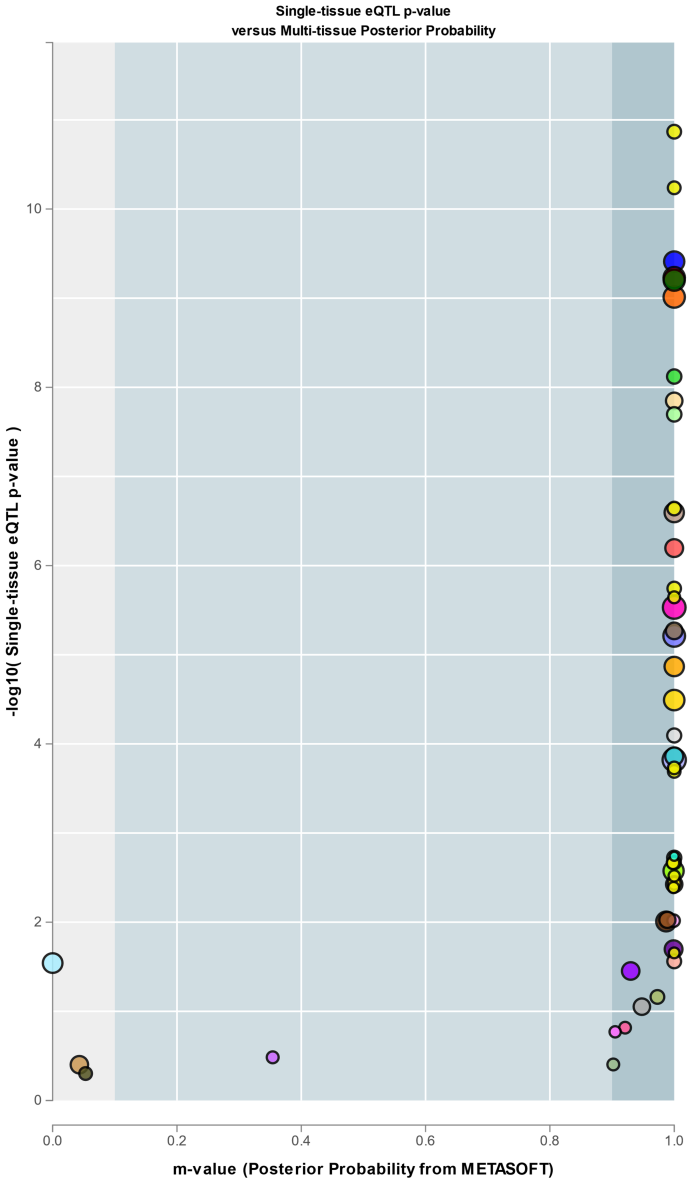


Figure 6.5.
Expression quantitative trait loci (eQTL) analysis of rs11062901 for *PARP11* expression from the GTEx database. Table of *P*-values for association of the SNP and *PARP11* expression in 49 different tissues. Significant tissues are highlighted in blue. The normalised effect size (NES) is defined as the slope of the linear regression and is computed as the effect of the alternative allele (T) relative to the reference allele (C). m-value (indicating the posterior probability that the effect is shared in each tissue tested in the cross-tissue meta-analysis, calculated by METASOFT) is plotted against -$\log_{10}(P)$ for each tissue.

6.3.4 Transcriptome-wide analysis

Gene expression levels were successfully predicted for 5,615 genes in whole-blood tissue and tested for an association with OS in patients with all wild-type CRC. The most significant gene was *MAP4K4* (HR=2.5x10^{34}, 95% CI=1.6x10^{19}-3.6x10^{49}, *P*=8.91x10^{-6}; **Figure 6.6**) although it did not pass the Bonferroni-corrected threshold for genome wide significance (*P*<8.9x10^{-6}) and is likely a statistical anomaly due to only 4 patients analysed having a non-zero GReX.

PARP11 was the second most strongly associated gene with OS (HR=0.093, 95% CI=0.03-0.26, *P*=1.08x10^{-5}). A reduction in *PARP11* GReX of 0.23 reduced median OS from 639 days to 421 days (**Figure 6.7**). Two eQTLs were annotated to PARP11 for imputation of expression levels.

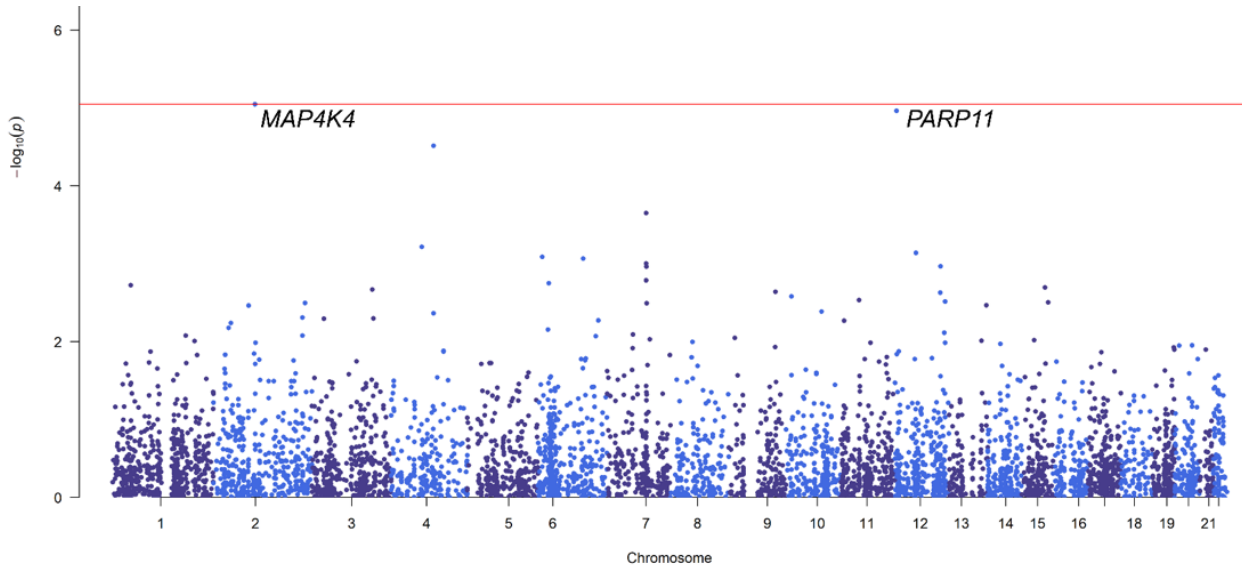


Figure 6.6. Manhattan plot of associations between predicted gene expression levels in whole-blood tissue and overall survival (OS) in 581 patients with wild-type colorectal cancers. Genes are ordered by chromosome position and plotted against the -log$_{10}$(*P*) for their association with OS. The red line represents the threshold for significance (*P*=8.9x10^{-6} based on a Bonferroni correction for 5,615 independent tests).

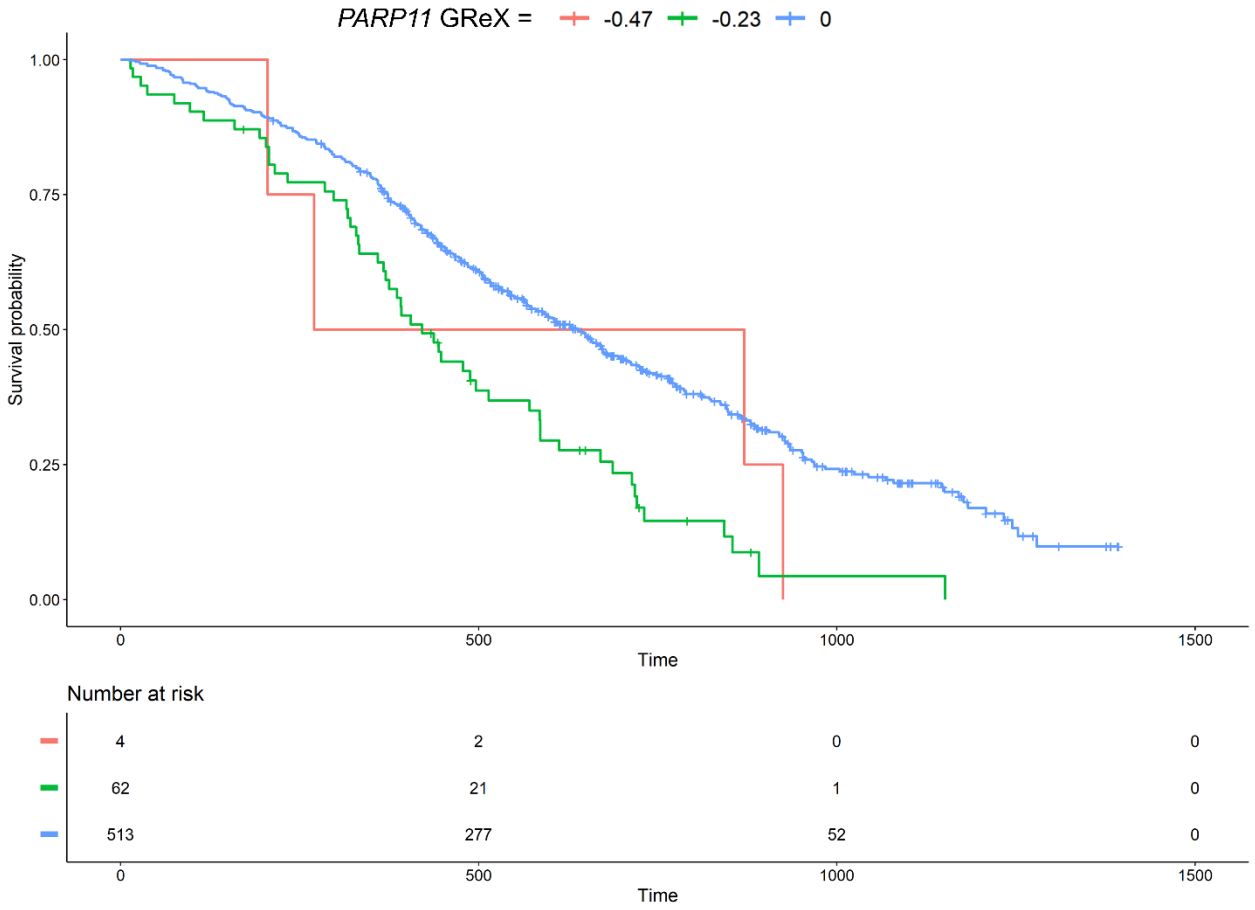


Figure 6.7. Kaplan-Meier plot for the relationship of genetically regulated gene expression (GReX) of *PARP11* in whole-blood tissue and overall survival in 579 patients with wild-type colorectal cancer. Time in days is plotted against survival probability for the 3 varying levels of GReX. The number of patients still at risk at each time point is shown beneath.

6.3.5 Analysis of *PARP11* expression and survival in THPA

PARP11 expression in tumour tissue was not associated with survival time in 597 unrelated patients (124 events) with colorectal tumours from THPA (*P*=0.14), nor in a subset with stage IV disease (n=83, events=39, *P*=0.63).

6.4 Discussion

6.4.1 Unmasking of a novel locus associated with survival

In my previous genome-wide analysis of OS in 1,926 unstratified patients from COIN and COIN-B, rs11062901 was not suggestive of association with survival in the SNP analyses (all COIN/COIN-B *P*=0.035, HR=1.17, 95% CI=1.01-1.36), and *PARP11* was not identified in the MAGMA gene level analyses (q=0.47; Chapter 3, Section 3.3). By excluding patients with CRCs carrying known somatic prognostic biomarkers, I have now shown that rs11062901 in *PARP11* and *PARP11* itself have a genome wide significant effect on survival. These data suggest that by excluding the known somatic biomarkers, I have effectively unmasked new genetic loci affecting survival. I have also started to understand the underlying mechanism. rs11062901 is associated with expression of *PARP11* in numerous tissues and *PARP11* expression itself was just under the threshold for genome-wide significance for survival in my TWAS. These data strongly suggest that decreased *PARP11* expression directly impairs survival outcomes. Data from the THPA suggests that this is not specifically due to expression levels in the colorectum and suggests a more general non-tissue specific mechanism.

6.4.2 *PARP11* expression and the tumour microenvironment

The tumour microenvironment (TME) has been shown to have an immunosuppressive effect; tumour cells can avoid normal immunosurveillance by manipulation of cytokines and the reprogramming of immune cells, allowing for progression of CRC and other cancers. Regulatory T (T_{reg}) cells, myeloid derived suppressor cells, cancer associated

fibroblasts, mast cells and tumour associated macrophages all create hostile conditions for the tumoricidal immune responses, including recruitment of $CD8^+$ cytotoxic T lymphocytes (CTLs) (Zhang et al. 2020). T_{reg} cells release adenosine which increases expression and hyperactivation of *PARP11* in CTLs, which aids in the ubiquitination and degradation of IFNAR1, without which the normal immune response is hindered. *PARP11* ablation in mice prevented loss of IFNAR1 and inhibited tumour growth due to increased CTL tumoricidal activity (Zhang et al. 2022). Therefore, I would expect that reduced *PARP11* expression would lead to improved prognosis in CRC patients by reducing IFNAR1 degradation.

In contrast to this, I have shown that germline variants that are predictive of reduced *PARP11* expression in whole blood show a detrimental effect on prognosis. This could be due to the tissue used in the TWAS; tumoricidal immune activity would be localised to the tumour or metastatic sites, not whole-blood. A cross-tissue TWAS analysis (Chapter 1, Section 1.3.1) could help elucidate on the organism-wide effects of *PARP11* GReX on CRC survival. However, this would still be susceptible to the limitations of imputing gene expression using germline variation, including potential bias in reference eQTL panels, and so direct measurement of expression in tumour tissue by RNA-sequencing would be more appropriate. No filtering for the number of annotated SNPs per gene was performed when selecting genes for the association analysis and so multiple genes had expression imputation based on the effects of a single eQTL. However, PARP11 used 2 SNPs in its imputation and the use of the singular most significant eQTL from the expression

reference panel is considered sufficient in many TWAS methodologies, despite losing some statistical power (Cao et al. 2022; Oliver et al. 2022).

6.4.3 Independent loci that passed the threshold for suggestive significance

The second most significant independent locus at 10p14 was not mapped to any coding genes and the sentinel SNP rs11254422 was not an eQTL in the GTeX database. However, the SNP does lie approximately 69Kb upstream of the oncogene *LINC00707* which has been shown to be upregulated in CRC tissue. High expression levels of *LINC00707* are predictive of advanced tumour stage, large size, distant metastasis, lymphatic metastasis, and poorer survival (Shao et al. 2019; Zhu et al. 2019; Wang et al. 2020). rs11254422 is also 172Kb downstream of the gene Protein Kinase C Theta (*PRKCQ*) and approximately 167Kb upstream of PRKCQ Antisense RNA 1 (*PRKCQ-AS1*). *PRKCQ* encodes PKC theta, a serine/threonine kinase that has been shown to promote growth, anoikis resistance, EMT and invasion in triple-negative breast cancer (Byerly et al. 2020). The long non-coding RNA *PRKCQ-AS1* has been shown to be overexpressed in CRC tissue and associated with poorer prognosis, possibly via mediation of the miR-1287-5p/YBX1 pathway (Cui et al. 2020).

I believe the association of SNPs with OS at this locus, which contains several oncogenes, makes it an interesting candidate for further study. However, relying on physical proximity alone can be a poor method for identifying causal genes. eQTL studies have suggested that two-thirds of the causal genes at significant GWAS loci are not the closest (Brænne et al. 2015; Zhu et al. 2016). Further QTL annotation for the lead SNPs

could include DNA methylation, protein expression, chromatin acetylation/chromatin accessibility and exon splicing. I have also previously discussed possible mechanisms of SNP effects on more distant genes (Chapter 4, Section 4.4.1). Differential expression analyses of the genes at this locus in samples from the COIN and COIN-B cohort could also find any potential associations with the SNPs of interest and CRC prognosis.

Chapter 7: General discussion

7.1 Novel findings and implications from my work

7.1.1 Germline prognostic biomarkers

I aimed to identify novel germline biomarkers of CRC survival to aid in patient care and management. Prior to this study, only a single variant in *CDH1* has been robustly validated as a prognostic germline biomarker despite many GWAS studies of CRC survival time (Chapter 1, Section 1.1.4.3). It is possible that this is due to the heterogeneity observed in CRC; many clinicopathological and somatic factors have prognostic effects that potentially eclipse the role of germline variants with smaller effect sizes. For this GWAS analysis, I have analysed the deeply phenotyped COIN and COIN-B cohorts for many of the established prognostic factors and, where possible, adjusted the regression analyses for those that showed a significant association with OS. In doing so, germline biomarkers of small effect may show an association with OS without the confounding effects of other factors, such as tumour surface area and resection status.

Although no variants reached strict genome-wide significance in the unstratified GWAS of all COIN and COIN-B patients, rs79612564 intronic to *ERBB4* was of suggestive significance (Chapter 3, Section 3.3.3). The minor (C) allele occurs in approximately 30% of Europeans and I showed it to be associated with a decrease in life expectancy of mCRC patients, with supporting mechanistic data. This finding was then nominally validated in mCRC patients from SOCCS and reached genome-wide significance when

meta-analysed with COIN and COIN-B. *ERBB4* is one of four members of the EGFR subfamily, which can heterodimerize with EGFR and activate downstream pathways such as PI3K-AKT-mTOR and MAPK/ERK (Lee et al. 2002).

7.1.2 Anatomy-specific germline biomarkers

It could be that the previous lack of evidence for germline prognostic biomarkers is due to the grouping of CRC cohorts for higher-powered analyses. By sub-grouping patient samples by primary tumour location, Labadie *et al.* (2022) observed site-specific germline variants associated with CRC survival. I replicated the effect of rs144717887 at 14q31.3 as a prognostic marker for proximal colon CRCs. I also identified the gene *PI4K2B* as significantly associated with OS in distal colon cancers specifically (Chapter 4, Section 4.3.3). The minor allele of the most significant variant mapped to *PI4K2B* was predictive of higher PI4K2B expression, which is associated with poorer survival in a separate cohort. Overall, these findings support the hypothesis that due to the differing embryological origins of gut tissues there may be tumour site-specific germline variation that is predictive of survival for CRC. Further studies should consider this when designing analyses.

7.1.3 Germline variation could identify treatment targets in difficult to treat cancers

MAPK-activated CRCs are difficult to treat due to their resistance to anti-EGFR antibody therapies (Chapter 2, Section 1.1.3.5). I aimed to find germline variation predictive of

survival in patients with these CRCs as a marker of potential treatment targets. In gene-level analysis of patients with MAPK-activated CRCs, *RASAL2* was the most strongly associated gene with OS, specifically in those with *KRAS*-mutant cancers. *RASAL2* directly interacts with RAS and so represents a strong candidate gene and potential therapy target. Upregulating *RASAL2* could enhance its GTPase activity converting RAS GTP to its inactive form.

7.1.4 Germline biomarkers in patients with CRCs without somatic prognostic mutations

Somatic mutations have considerable effects on disease progression and prognosis (Chapter 1, Section 1.1.4.2). By removing patients with known somatic prognostic biomarkers from the GWAS analysis I hoped to further remove any confounding effects on prognosis and identify germline markers of smaller effect size. A significant association between OS and *PARP11* was observed. This gene was not significant under any of the previous analyses, supporting the hypothesis that prognostic germline alleles can be detected on a cleaner somatic background.

PARP11 remains a poorly studied gene in the context of CRC. However, one study observed that ablation of *PARP11* hindered tumour growth in a mouse model via regulation of the TME (Zhang et al. 2022). This contrasts with the TWAS analysis presented here; reduced *PARP11* expression was strongly, but not significantly, associated with poorer survival in whole-blood tissue. One explanation for this could be the tissue specificity of expression-based analyses.

7.2 Strengths and limitations

7.2.1 Validation cohorts

Due to the lack of a significant difference in survival time between treatment arms in COIN and COIN-B (Chapter 2, Section 2.3.1) I was able to combine all patient groups into a relatively large clinical cohort with a wealth of clinicopathological and somatic data available for analysis. However, the gold-standard for biomarker discovery remains replication of any statistical associations in external patient cohorts to ensure they are not chance findings (Kraft et al. 2009). Unfortunately, I was unable to find suitable validation cohorts to properly replicate the associations in chapters 5 and 6, as few clinical studies collected the necessary somatic data. As such, the SNP associations with survival could be chance findings unique to the COIN and COIN-B cohorts. I was able to replicate anatomy specific variation observed in COIN and COIN-B using the UK Biobank, as well as one of the findings from Labadie *et al.* (2022) in proximal colon tumours. In Chapter 3, rs79612564 (2q34, intronic to *ERBB4*) nominally validated in mCRC patients from SOCCS, but not from ISACC. This is possibly due to the confounding effects of other clinical and pathological factors that could not be adjusted for in the population-based studies.

7.2.2 "I (may not) Have the Power!"

In line with recommended GWAS QC measures and the sample size of this study, MAF≥0.05 was set for inclusion of SNPs in all GWAS. At this threshold I only had sufficient (>80%) power to detect genome-wide significant SNP associations with a HR>1.69 under an additive model in the 1,926-patient cohort. The 493 patients with distal colon cancers represent the smallest stratified sample and had an equivalent detectable HR>2.78. Despite these GWAS analyses being some of the largest of their kind in mCRC, these effect sizes are still unlikely to be observed in common germline variant analysis. For example, the only robustly validated germline biomarker of survival, rs9929218 at 16q22, only had a HR of 1.28 (95% CI=1.14-1.43) in the combined analysis of training and validation cohorts (Smith et al. 2015). Of all the 205 CRC risk SNPs outlined in Fernadez-Rozadilla *et al.* (2023), only a single variant, rs201395236 at 1q44 (Lu et al. 2019), had an observed beta coefficient greater than that detectable in my largest analysis (beta=-0.528, equivalent in magnitude of effect to HR=1.70).

Multiple testing correction was observed throughout this work. The most used method was Bonferroni correction (Armstrong 2014) as it is the *de facto* standard for many of the analyses performed, including the genome-wide significance threshold (Chapter 1, Section 1.2.2.3). However, Bonferroni correction is considered overly conservative in many cases (Perneger 1998), possibly increasing the false-negative rate. A less conservative FDR adjustment of *P*-values may be more appropriate for many of these analyses (Benjamini and Hochberg 1995), such as gene-level MAGMA analysis. There is also the 'winner's curse' (Bazerman and Samuelson 1983) to consider. This describes

the phenomenon where estimators of association and effect size for significant findings are often upwardly biased in discovery cohorts, leading to ascertainment bias. If effect sizes are initially overestimated, then follow up studies will be underpowered and fail. Therefore I may not have had sufficient power to replicate the true effect size in the external cohorts available (Xiao and Boehnke 2009).

7.2.3 From variation to causation

The main aim of this study was to identify germline variants that could predict patient prognosis. However, of equal importance is deciphering the exact biological mechanisms by which these genetic variants have an effect and therefore better understand CRC disease progression. This can prove difficult, as significant GWAS hits are likely capturing the effect of causal variants due to LD rather than being the causal variants themselves, misleading downstream mechanistic analyses (Uffelmann et al. 2021). Also, the hits are most often intergenic, sometimes intronic and rarely protein coding, making their interpretation more difficult.

MAGMA gene-based analysis (de Leeuw et al. 2015) allows for individual SNP associations to be annotated to genes by chromosomal position and their cumulative association used to test for the association of genes with the phenotype of interest. For this study the SNP annotation window was set to 35Kb upstream of the gene transcription zone and 10Kb downstream, based upon examples from the current literature (Sey et al. 2020; Liu et al. 2021). This is to capture variation in the promoter regions of genes and any other cis regulatory elements that could potentially affect gene expression. However, there are no universally agreed values for this window and there is evidence that

variations in window size can have large effects on the number of significant associations, despite not affecting power (de Leeuw et al. 2015). In future it may be important to study the effect of varying the window size on any significant findings.

In Chapter 3 I used MAGMA version v1.07. It was later found by Yurko *et al.* (2021) that this version had an inflated false-positive rate, especially for larger genes, due to its implementation of Brown's approximation of Fisher's method for combining dependent SNP-level *P*-values to adjust for their LD-induced covariance. In response, Leeuw *et al.* (2020) amended the SNP-wise mean model in MAGMA v1.08. However, no significant gene or gene-set associations were reported in the Chapter 3 analyses using the earlier version of MAGMA, making the inflated false-positive rate unimpactful upon this study.

eQTL and the transcriptome-wide analyses made use of the GTEx reference dataset (Chapter 2, Section 2.3.5) to find associations between candidate SNPs and gene expression, elucidating on causal mechanisms of SNP effect. A causal SNP that is also an eQTL could be falsely capturing the effect of another eQTL due to LD and so is a false positive mechanistic finding. Colocalization analysis, using software tools such as eCAVIAR (Hormozdiari et al. 2016) and HyPrColoc (Foley et al. 2021) determines whether a single SNP is responsible for both the eQTL and GWAS signals. This could improve the reliability of some causal inferences made in this study, such as the A allele of rs313566 potentially increasing the expression of *PI4K2B* and thus improving prognosis in patients with distal colon tumours (Chapter 4, Section 4.3.3).

Expression analyses are highly tissue specific, with some variants having inverse effects in different cell types (Mizuno and Okada 2019). This has made interpretation of the identified eQTLs difficult as mCRC is an extremely heterogeneous disease that affects many tissues throughout the body outside of the colon. Whole blood expression panels are often used in TWAS analyses (Wainberg et al. 2019), as seen in Chapter 6. This is to maximise power as whole blood is the second most analysed tissue in the GTEx dataset after skeletal muscle (n=755 and 803, respectively). Also, whole blood is considered a suitable surrogate when there are no clear candidate tissues of interest due to its sharing of >80% of the transcriptome with colon, brain, heart, kidney, liver, lung, prostate, spleen and stomach tissue (Liew et al. 2006; Mehta et al. 2013). However, in this study the surrogate tissue has not assisted in narrowing down the true biological mechanisms and tissues in which the expression of these genes is having an effect. Therefore, further individual TWAS analyses in other candidate tissues are warranted, or a multi-tissue approach (Chapter 1, Section 1.3.1), preferably using direct RNA-sequencing information instead of imputed GReX levels. One such multi-tissue approach is UTMOST (Unified Test for MOlecular SignaTures; https://github.com/Joker-Jerome/UTMOST), a statistical framework for producing cross-tissue expression imputation and gene-level association analysis (Hu et al. 2019).

7.2.4 Clinical utility

Only a select few somatic genetic markers of CRC prognosis are routinely used by clinicians. Examples include *BRAF* V600E and *KRAS* mutations due to their effect sizes and effect on treatment options (Chapter 1, Section 1.1.4.2). The clinical utility of germline

variants with smaller effect sizes as standalone markers is very low. However, like many phenotypes studied by GWAS (Visscher et al. 2017; Uffelmann et al. 2021), CRC prognosis could prove to be highly polygenic, making the cumulative effect of many germline associations an effective predictive tool. Evidence for this comes from the gene-sets significantly associated with OS presented here, such as 'Negative regulation of phospholipid biosynthetic process' in rectal cancers (Chapter 4, Section 4.3.4) and 'Golgi cisterna membrane' in MAPK-activated cancers (Chapter 5, Section 5.3.8). By annotating SNPs to genes and then genes to gene-sets I tested the cumulative association of these SNPs across large sections of the genome. Therefore, the significant association between OS and these gene-sets may suggest a polygenic model of inheritance.

Polygenic risk scores (PRS) allow us to use GWAS summary statistics to quantify the cumulative effect of SNP variation across the genome on a trait of interest, such as CRC prognosis. PRS are calculated by multiplying the count of DNA variants with predetermined trait-specific effect sizes and provide useful predictive models of an individual's genetic susceptibility to a trait (Wray et al. 2021). SNPs are most often selected by assigning a threshold for significance from a discovery GWAS, adjusting this threshold to maximise the PRS specificity and sensitivity in a training dataset and then testing its validity in a validation cohort. The number of included SNPs can vary greatly by *P*-value threshold, but there are also effect size (reported as odds ratio or beta coefficient) shrinkage techniques that allow for inclusion of all SNPs from the discovery cohort regardless of association (Choi et al. 2020). In a clinical setting it is likely that a PRS would be calculated via a custom genotyping panel containing all the SNPs of

interest. Any strongly associated SNPs with relevant validation, such as those presented in this study, could be included in this SNP panel, or imputed in separately.

Despite no PRS existing for CRC prognosis due to the low number of significantly associated loci, PRS models have been extensively tested for CRC risk. Sassano *et al.* (2022) reviewed 33 independent studies and found that the addition of these genetic factors to models containing traditional risk-factors enhanced the area under the curve (AUC) values by an average of 0.040 (range 0.010-0.084), although most could still not reach the preferred threshold for discriminatory accuracy (AUC>0.70) (Swets 1988). The models also had heterogeneity in their methodology (some used unweighted allele counts) and size (4-696 SNPs included). It was found that including a greater number of SNPs in the models did not improve the model's predictive accuracy.

The predictive power of PRS is limited to the contribution of common genetic variation on the trait and ignores the potentially large effects of environmental factors and rare variants undiscoverable by traditional GWAS methods (Wray et al. 2021). Current estimates of the typical PRS sensitivity for disease risk prediction are 10-15% when specificity is set to 95%. That is, when the number of people with high PRS not developing the disease is reduced to below 5% the PRS will accurately predict 10-15% of people who will go on to develop the disease (Sud et al. 2023). There is also debate of the clinical validity of PRS versus their clinical utility. In a recent systematic review of PRS it was found that many studies demonstrated their effectiveness for disease prediction (clinical validity) but none were able to show an unequivocal improvement for patient outcomes (clinical utility)

(Kumuthini et al. 2022). If the clinical utility rates do not improve, especially considering the economic cost of screening, it is unlikely that we will see their widespread use anytime soon.

7.2.5 Transferability and ethics

To reduce the effects of population stratification on false-positive rates it is necessary to reduce GWAS populations to a single genetic ancestry. Because of differing LD structures and allele frequencies, germline variants identified by GWAS that are not robustly verified as causal cannot be generalised to genetic ancestries outside of these studied populations (Carlson et al. 2013; Uffelmann et al. 2021). This especially applies to PRS; a recent study of PRS across populations found that their predictive accuracy is inversely proportional to the Euclidian distance of genetic principle components for the target population from those of the discovery cohort (Ding et al. 2023). Due to the availability of data, most GWAS studies use individuals of European ancestry leaving other populations severely understudied, particularly those of low socio-economic status. This reduces the global clinical utility of GWAS findings and leads to ethical concerns around diversity and inclusion, as these individuals cannot receive the health benefits. As researchers we should be working to make our outputs more generalisable and future study could include other diverse genetic ancestries.

7.3 Future work

Although the imputation quality of rs79612564 had a >99% concordance with the independent KASPar genotyping (Chapter 3, Section 3.3.3) it may be important to confirm the genotyping accuracy for the other SNP biomarkers presented in this thesis, especially those with a lower imputation quality score.

RNA-sequencing of the COIN and COIN-B tumour samples could allow for more reliable eQTL and transcriptome-wide survival analyses and act as replication for THPA findings presented throughout this thesis. Differential expression analysis between healthy and disease tissues could identify dysregulated genes and pathways in CRC tumours. Similarly, a methylome-wide association study would enable the integration of DNA methylation reference datasets with the COIN and COIN-B SNP genotyping to study the effects of epigenetic regulation on CRC prognosis. This has already been used to identify novel loci associated with CRC risk (Fernandez-Rozadilla et al. 2023).

Replication of the prognostic biomarkers presented in Chapters 5 and 6 is vital for their utility. As somatic mutation testing becomes more prevalent in the clinic and medical records are linked to biobank size datasets, it may become viable to form suitable validation cohorts of MAPK-activated and wild-type mCRC patients. Wet lab-based techniques could also be used to test the validity of the candidate therapeutic targets identified here, such as *RASAL2*. One study found that *RASAL2* ablation in a mouse model of luminal B breast cancer resulted in enhanced metastasis via upregulation of MEK/ERK and PI3K/AKT signalling (Olsen et al. 2017). A similar study of *RASAL2*

upregulation in a CRC mouse model or cell lines with activating-*KRAS* mutations could help confirm a similar mechanism in MAPK-activated CRCs and therefore its relevance as a therapeutic target.

7.4 Outlook

The work in this thesis has identified novel germline prognostic biomarkers for mCRC patients by tumour location and somatic mutation status, as well as potential therapeutic targets. While many of these SNPs and genes have relatively small effect sizes and have not yet been robustly validated in external replication cohorts due to lack of available data, their inclusion in polygenic models of CRC prognosis could be of clinical utility in the future.

References

Abuli, A. et al. 2013. Genetic susceptibility variants associated with colorectal cancer prognosis. *Carcinogenesis* 34(10), pp. 2286-2291. doi: 10.1093/carcin/bgt179

Adams, R. A. et al. 2011. Intermittent versus continuous oxaliplatin and fluoropyrimidine combination chemotherapy for first-line treatment of advanced colorectal cancer: results of the randomised phase 3 MRC COIN trial. *Lancet Oncology* 12(7), pp. 642-653. doi: 10.1016/s1470-2045(11)70102-4

Afolabi, H. A. et al. 2022. A GNAS Gene Mutation's Independent Expression in the Growth of Colorectal Cancer: A Systematic Review and Meta-Analysis. *Cancers (Basel)* 14(22), doi: 10.3390/cancers14225480

Al-Tassan, N. et al. 2002. Inherited variants of MYH associated with somatic G:C→T:A mutations in colorectal tumors. *Nature Genetics* 30(2), pp. 227-232. doi: 10.1038/ng828

Al-Tassan, N. A. et al. 2015. A new GWAS and meta-analysis with 1000Genomes imputation identifies novel risk variants for colorectal cancer. *Sci Rep* 5, p. 10442. doi: 10.1038/srep10442

Aleksandrova, K. et al. 2014. Combined impact of healthy lifestyle factors on colorectal cancer: a large European cohort study. *BMC Med* 12, p. 168. doi: 10.1186/s12916-014-0168-4

Algra, A. M. and Rothwell, P. M. 2012. Effects of regular aspirin on long-term cancer incidence and metastasis: a systematic comparison of evidence from observational studies versus randomised trials. *Lancet Oncol* 13(5), pp. 518-527. doi: 10.1016/s1470-2045(12)70112-2

References

Alhopuro, P. et al. 2005. SMAD4 levels and response to 5-fluorouracil in colorectal cancer. *Clin Cancer Res* 11(17), pp. 6311-6316. doi: 10.1158/1078-0432.ccr-05-0244

Allegra, C. J. et al. 2009. American Society of Clinical Oncology Provisional Clinical Opinion: Testing for KRAS Gene Mutations in Patients With Metastatic Colorectal Carcinoma to Predict Response to Anti-Epidermal Growth Factor Receptor Monoclonal Antibody Therapy. *Journal of Clinical Oncology* 27(12), pp. 2091-2096. doi: 10.1200/jco.2009.21.9170

Alli-Baloguna, G. O., Gewinner, C. A., Jacobs, R., Kriston-Vizi, J., Waugh, M. G. and Minogue, S. 2016. Phosphatidylinositol 4-kinase II beta negatively regulates invadopodia formation and suppresses an invasive cellular phenotype. *Molecular Biology of the Cell* 27(25), pp. 4033-4042. doi: 10.1091/mbc.E16-08-0564

Altshuler, D., Donnelly, P. and The International HapMap, C. 2005. A haplotype map of the human genome. *Nature* 437(7063), pp. 1299-1320. doi: 10.1038/nature04226

Altshuler, D. M. et al. 2015. A global reference for human genetic variation. *Nature* 526(7571), pp. 68-+. doi: 10.1038/nature15393

Amin, M. B. et al. 2017. The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. *CA Cancer J Clin* 67(2), pp. 93-99. doi: 10.3322/caac.21388

Anaconda_inc. 2020. *Anaconda Software Distribution*. Anaconda Inc.

Anderson, G. et al. 1998. Design of the Women's Health Initiative Clinical Trial and Observational Study. *Controlled Clinical Trials* 19(1), pp. 61-109.

References

Andreyev, H. J. N., Norman, A. R., Cunningham, D., Oates, J. R., Clarke, P. A. and Grp, R. 1998. Kirsten ras mutations in patients with colorectal cancer: the multicenter "RASCAL" study. *Journal of the National Cancer Institute* 90(9), pp. 675-684. doi: 10.1093/jnci/90.9.675

Arem, H., Moore, S. C., Park, Y., Ballard-Barbash, R., Hollenbeck, A., Leitzmann, M. and Matthews, C. E. 2014. Physical activity and cancer-specific mortality in the NIH-AARP Diet and Health Study cohort. *Int J Cancer* 135(2), pp. 423-431. doi: 10.1002/ijc.28659

Armstrong, R. A. 2014. When to use the Bonferroni correction. *Ophthalmic Physiol Opt* 34(5), pp. 502-508. doi: 10.1111/opo.12131

Ashburner, M. et al. 2000. Gene Ontology: tool for the unification of biology. *Nature Genetics* 25(1), pp. 25-29. doi: 10.1038/75556

Astle, W. J. et al. 2016. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell* 167(5), pp. 1415-1429.e1419. doi: https://doi.org/10.1016/j.cell.2016.10.042

Barault, L. et al. 2008. Hypermethylator Phenotype in Sporadic Colon Cancer: Study on a Population-Based Series of 582 Cases. *Cancer Research* 68(20), pp. 8541-8546. doi: 10.1158/0008-5472.CAN-08-1171

Barbeira, A. N. et al. 2021. Exploiting the GTEx resources to decipher the mechanisms at GWAS loci. *Genome Biology* 22(1), p. 49. doi: 10.1186/s13059-020-02252-4

Barbeira, A. N., Pividori, M., Zheng, J., Wheeler, H. E., Nicolae, D. L. and Im, H. K. 2019. Integrating predicted transcriptome from multiple tissues improves association detection. *PLoS Genet* 15(1), p. e1007889. doi: 10.1371/journal.pgen.1007889

References

Bardou, M., Barkun, A. and Martel, M. 2010. Effect of statin therapy on colorectal cancer. *Gut* 59(11), pp. 1572-1585. doi: 10.1136/gut.2009.190900

Bazerman, M. H. and Samuelson, W. F. 1983. I Won the Auction But Don't Want the Prize. *Journal of Conflict Resolution* 27(4), pp. 618-634. doi: 10.1177/0022002783027004003

Belanger, C., Speizer, F. E., Hennekens, C. H., Rosner, B., Willett, W. and Bain, C. 1980. THE NURSES HEALTH STUDY - CURRENT FINDINGS. *American Journal of Nursing* 80(7), pp. 1333-1333. doi: 10.1097/00000446-198007000-00024

Benjamini, Y. and Hochberg, Y. 1995. Controlling the false discovery rate - a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B-Statistical Methodology* 57(1), pp. 289-300.

Bernards, A. 2003. GAPs galore! A survey of putative Ras superfamily GTPase activating proteins in man and Drosophila. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer* 1603(2), pp. 47-82. doi: https://doi.org/10.1016/S0304-419X(02)00082-3

Bingmer, K. et al. 2020. Primary tumor location impacts survival in colorectal cancer patients after resection of liver metastases. *J Surg Oncol*, doi: 10.1002/jso.26061

Botteri, E., Iodice, S., Bagnardi, V., Raimondi, S., Lowenfels, A. B. and Maisonneuve, P. 2008. Smoking and colorectal cancer: a meta-analysis. *Jama* 300(23), pp. 2765-2778. doi: 10.1001/jama.2008.839

Brenner, H., Chang-Claude, J., Rickert, A., Seiler, C. M. and Hoffmeister, M. 2012. Risk of Colorectal Cancer After Detection and Removal of Adenomas at Colonoscopy: Population-Based Case-Control Study. *Journal of Clinical Oncology* 30(24), pp. 2969-2976. doi: 10.1200/jco.2011.41.3377

References

Brenner, H., Chang-Claude, J., Seiler, C. M., Rickert, A. and Hoffmeister, M. 2011. Protection From Colorectal Cancer After Colonoscopy A Population-Based, Case-Control Study. *Annals of Internal Medicine* 154(1), pp. 22-U156. doi: 10.7326/0003-4819-154-1-201101040-00004

Brenner, H., Kloor, M. and Pox, C. P. 2014. Colorectal cancer. *Lancet* 383(9927), pp. 1490-1502. doi: 10.1016/s0140-6736(13)61649-9

Brodie, A., Azaria, J. R. and Ofran, Y. 2016. How far from the SNP may the causative genes be? *Nucleic acids research* 44(13), pp. 6046-6054. doi: 10.1093/nar/gkw500

Brænne, I. et al. 2015. Prediction of Causal Candidate Genes in Coronary Artery Disease Loci. *Arterioscler Thromb Vasc Biol* 35(10), pp. 2207-2217. doi: 10.1161/atvbaha.115.306108

Bufill, J. A. 1990. Colorectal-cancer - evidence for distinct genetic categories based on proximal or distal tumor location. *Annals of Internal Medicine* 113(10), pp. 779-788. doi: 10.7326/0003-4819-113-10-779

Bui, S., Mejia, I., Díaz, B. and Wang, Y. 2021. Adaptation of the Golgi Apparatus in Cancer Cell Invasion and Metastasis. *Frontiers in Cell and Developmental Biology* 9,

Bycroft, C. et al. 2018. The UK Biobank resource with deep phenotyping and genomic data. *Nature* 562(7726), pp. 203-209. doi: 10.1038/s41586-018-0579-z

Byerly, J. H., Port, E. R. and Irie, H. Y. 2020. PRKCQ inhibition enhances chemosensitivity of triple-negative breast cancer by regulating Bim. *Breast Cancer Research* 22(1), p. 10. doi: 10.1186/s13058-020-01302-w

References

Calle, E. E. et al. 2002. The American Cancer Society cancer prevention study II nutrition cohort - Rationale, study design, and baseline characteristics. *Cancer* 94(9), pp. 2490-2501. doi: 10.1002/cncr.101970

Cancer Genome Atlas Research, N. et al. 2013. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature genetics* 45(10), pp. 1113-1120. doi: 10.1038/ng.2764

CancerResearchUK. 2023. *Bowel cancer statistics*. Available at: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/bowel-cancer#heading-Zero [Accessed: 10/04/2023].

Cao, C. et al. 2022. webTWAS: a resource for disease candidate susceptibility genes identified by transcriptome-wide association study. *Nucleic Acids Res* 50(D1), pp. D1123-d1130. doi: 10.1093/nar/gkab957

Carithers, L. J. and Moore, H. M. 2015. The Genotype-Tissue Expression (GTEx) Project. *Biopreservation and Biobanking* 13(5), pp. 307-308. doi: 10.1089/bio.2015.29031.hmm

Carlson, C. S. et al. 2013. Generalization and Dilution of Association Results from European GWAS in Populations of Non-European Ancestry: The PAGE Study. *PLOS Biology* 11(9), p. e1001661. doi: 10.1371/journal.pbio.1001661

Cathomas, G. 2014. PIK3CA in Colorectal Cancer. *Front Oncol* 4, p. 35. doi: 10.3389/fonc.2014.00035

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, p. 7. doi: 10.1186/s13742-015-0047-8

References

Chen, H.-S. and Sheen-Chen, S.-M. 2000. Obstruction and perforation in colorectal adenocarcinoma: An analysis of prognosis and current trends. *Surgery* 127(4), pp. 370-376. doi: https://doi.org/10.1067/msy.2000.104674

Chen, L., Ye, L. and Hu, B. 2022. Hereditary Colorectal Cancer Syndromes: Molecular Genetics and Precision Medicine. *Biomedicines* 10(12), doi: 10.3390/biomedicines10123207

Choi, S. W., Mak, T. S. and O'Reilly, P. F. 2020. Tutorial: a guide to performing polygenic risk score analyses. *Nat Protoc* 15(9), pp. 2759-2772. doi: 10.1038/s41596-020-0353-1

Chou, W.-C. et al. 2016. A combined reference panel from the 1000 Genomes and UK10K projects improved rare variant imputation in European and Chinese samples. *Scientific Reports* 6(1), p. 39313. doi: 10.1038/srep39313

Cohen, R. B. 2003. Epidermal growth factor receptor as a therapeutic target in colorectal cancer. *Clin Colorectal Cancer* 2(4), pp. 246-251. doi: 10.3816/CCC.2003.n.006

Colditz, G. A., Manson, J. E. and Hankinson, S. E. 1997. The Nurses' Health Study: 20-year contribution to the understanding of health among women. *Journal of Womens Health* 6(1), pp. 49-62. doi: 10.1089/jwh.1997.6.49

Cremolini, C. et al. 2015a. BRAF codons 594 and 596 mutations identify a new molecular subtype of metastatic colorectal cancer at favorable prognosis. *Ann Oncol* 26(10), pp. 2092-2097. doi: 10.1093/annonc/mdv290

Cremolini, C. et al. 2015b. FOLFOXIRI plus bevacizumab versus FOLFIRI plus bevacizumab as first-line treatment of patients with metastatic colorectal cancer: updated overall survival and molecular subgroup analyses of the open-label, phase 3

References

TRIBE study. *Lancet Oncology* 16(13), pp. 1306-1315. doi: 10.1016/s1470-2045(15)00122-9

Cui, G. C., Zhao, H. L. and Li, L. N. 2020. Long noncoding RNA PRKCQ-AS1 promotes CRC cell proliferation and migration via modulating miR-1287-5p/YBX1 axis. *Journal of Cellular Biochemistry* 121(10), pp. 4166-4175. doi: 10.1002/jcb.29712

Dahm, C. C. et al. 2010. Dietary Fiber and Colorectal Cancer Risk: A Nested Case–Control Study Using Food Diaries. *JNCI: Journal of the National Cancer Institute* 102(9), pp. 614-626. doi: 10.1093/jnci/djq092

Dai, J. et al. 2012. GWAS-identified colorectal cancer susceptibility loci associated with clinical outcomes. *Carcinogenesis* 33(7), pp. 1327-1331. doi: 10.1093/carcin/bgs147

Das, S., Abecasis, G. R. and Browning, B. L. 2018. Genotype Imputation from Large Reference Panels. *Annu Rev Genomics Hum Genet* 19, pp. 73-96. doi: 10.1146/annurev-genom-083117-021602

Dawson, H., Kirsch, R., Driman, D. K., Messenger, D. E., Assarzadegan, N. and Riddell, R. H. 2014. Optimizing the detection of venous invasion in colorectal cancer: the ontario, Canada, experience and beyond. *Front Oncol* 4, p. 354. doi: 10.3389/fonc.2014.00354

de Leeuw, C. A., Mooij, J. M., Heskes, T. and Posthuma, D. 2015. MAGMA: Generalized Gene-Set Analysis of GWAS Data. *Plos Computational Biology* 11(4), p. 19. doi: 10.1371/journal.pcbi.1004219

De Palma, F. D., D'Argenio, V., Pol, J., Kroemer, G., Maiuri, M. C. and Salvatore, F. 2019. The Molecular Hallmarks of the Serrated Pathway in Colorectal Cancer. *Cancers* 11 (7). doi: 10.3390/cancers11071017

References

De Roock, W. et al. 2010. Effects of KRAS, BRAF, NRAS, and PIK3CA mutations on the efficacy of cetuximab plus chemotherapy in chemotherapy-refractory metastatic colorectal cancer: a retrospective consortium analysis. *Lancet Oncol* 11(8), pp. 753-762. doi: 10.1016/s1470-2045(10)70130-3

Ding, Y. et al. 2023. Polygenic scoring accuracy varies across the genetic ancestry continuum. *Nature*, doi: 10.1038/s41586-023-06079-4

Dobrzynska, I., Szachowicz-Petelska, B., Sulkowski, S. and Figaszewski, Z. 2005. Changes in electric charge and phospholipids composition in human colorectal cancer cells. *Molecular and Cellular Biochemistry* 276(1-2), pp. 113-119. doi: 10.1007/s11010-005-3557-3

Dotor, E. et al. 2006. Tumor thymidylate synthase 1494del6 genotype as a prognostic factor in colorectal cancer patients receiving fluorouracil-based adjuvant treatment. *J Clin Oncol* 24(10), pp. 1603-1611. doi: 10.1200/jco.2005.03.5253

Doubeni, C. A. et al. 2012. Contribution of behavioral risk factors and obesity to socioeconomic differences in colorectal cancer incidence. *J Natl Cancer Inst* 104(18), pp. 1353-1362. doi: 10.1093/jnci/djs346

Douillard, J. Y. et al. 2010. Randomized, Phase III Trial of Panitumumab With Infusional Fluorouracil, Leucovorin, and Oxaliplatin (FOLFOX4) Versus FOLFOX4 Alone As First-Line Treatment in Patients With Previously Untreated Metastatic Colorectal Cancer: The PRIME Study. *Journal of Clinical Oncology* 28(31), pp. 4697-4705. doi: 10.1200/jco.2009.27.4860

Douillard, J. Y. et al. 2014. Final results from PRIME: randomized phase III study of panitumumab with FOLFOX4 for first-line treatment of metastatic colorectal cancer. *Ann Oncol* 25(7), pp. 1346-1355. doi: 10.1093/annonc/mdu141

References

Dowle, M. and Srinivasan, A. 2019. data.table: Extension of `data.frame`.

Ehret, G. B. 2010. Genome-wide association studies: contribution of genomics to understanding blood pressure and essential hypertension. *Curr Hypertens Rep* 12(1), pp. 17-25. doi: 10.1007/s11906-009-0086-6

Eklof, V. et al. 2013. The prognostic role of KRAS, BRAF, PIK3CA and PTEN in colorectal cancer. *British Journal of Cancer* 108(10), pp. 2153-2163. doi: 10.1038/bjc.2013.212

Erdrich, J., Zhang, X., Giovannucci, E. and Willett, W. 2015. Proportion of colon cancer attributable to lifestyle in a cohort of US women. *Cancer Causes Control* 26(9), pp. 1271-1279. doi: 10.1007/s10552-015-0619-z

Fang, J. F., Zhao, H. P., Wang, Z. F. and Zheng, S. S. 2017. Upregulation of RASAL2 promotes proliferation and metastasis, and is targeted by miR-203 in hepatocellular carcinoma. *Mol Med Rep* 15(5), pp. 2720-2726. doi: 10.3892/mmr.2017.6320

Fang, S. and Wang, Z. 2014. EGFR mutations as a prognostic and predictive marker in non-small-cell lung cancer. *Drug Des Devel Ther* 8, pp. 1595-1611. doi: 10.2147/dddt.s69690

Faron, M. et al. 2015. Is primary tumour resection associated with survival improvement in patients with colorectal cancer and unresectable synchronous metastases? A pooled analysis of individual data from four randomised trials. *Eur J Cancer* 51(2), pp. 166-176. doi: 10.1016/j.ejca.2014.10.023

Fearnhead, N. S., Britton, M. P. and Bodmer, W. F. 2001. The ABC of APC. *Hum Mol Genet* 10(7), pp. 721-733. doi: 10.1093/hmg/10.7.721

References

Fearon, E. R. 2011. Molecular genetics of colorectal cancer. *Annu Rev Pathol* 6, pp. 479-507. doi: 10.1146/annurev-pathol-011110-130235

Fedirko, V. et al. 2011. Alcohol drinking and colorectal cancer risk: an overall and dose–response meta-analysis of published studies. *Annals of Oncology* 22(9), pp. 1958-1972. doi: https://doi.org/10.1093/annonc/mdq653

Feng, M. et al. 2014. RASAL2 activates RAC1 to promote triple-negative breast cancer progression. *The Journal of clinical investigation* 124(12), pp. 5291-5304. doi: 10.1172/JCI76711

Fernandez-Rozadilla, C. et al. 2023. Deciphering colorectal cancer genetics through multi-omic analysis of 100,204 cases and 154,587 controls of European and east Asian ancestries. *Nature Genetics* 55(1), pp. 89-99. doi: 10.1038/s41588-022-01222-9

Foley, C. N., Staley, J. R., Breen, P. G., Sun, B. B., Kirk, P. D. W., Burgess, S. and Howson, J. M. M. 2021. A fast and efficient colocalization algorithm for identifying shared genetic risk factors across multiple traits. *Nature Communications* 12(1), p. 764. doi: 10.1038/s41467-020-20885-8

Foster, D. A. 2009. Phosphatidic acid signaling to mTOR: Signals for the survival of human cancer cells. *Biochimica Et Biophysica Acta-Molecular and Cell Biology of Lipids* 1791(9), pp. 949-955. doi: 10.1016/j.bbalip.2009.02.009

Galon, J. et al. 2006. Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. *Science* 313(5795), pp. 1960-1964. doi: 10.1126/science.1129139

Gamazon, E. R. et al. 2015a. A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet* 47(9), pp. 1091-1098. doi: 10.1038/ng.3367

References

Gamazon, E. R. et al. 2015b. A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics* 47(9), pp. 1091-+. doi: 10.1038/ng.3367

Garcia-Albeniz, X. et al. 2013. Phenotypic and tumor molecular characterization of colorectal cancer in relation to a susceptibility SMAD7 variant associated with survival. *Carcinogenesis* 34(2), pp. 292-298. doi: 10.1093/carcin/bgs335

Gentleman, R. C. et al. 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology* 5(10), p. R80. doi: 10.1186/gb-2004-5-10-r80

Giles, G. G. and English, D. R. 2002. The Melbourne Collaborative Cohort Study. *IARC Sci Publ* 156, pp. 69-70.

Gohagan, J. K., Prorok, P. C., Hayes, R. B., Kramer, B. S. and Team, P. P. 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), pp. 251S-272S. doi: 10.1016/s0197-2456(00)00097-0

Goldberg, R. M. et al. 2004. A Randomized controlled trial of fluorouracil plus leucovorin, irinotecan, and oxaliplatin combinations in patients with previously untreated metastatic colorectal cancer. *Journal of Clinical Oncology* 22(1), pp. 23-30. doi: 10.1200/jco.2004.09.046

Guan, W. L. et al. 2020. Clinicopathologic Features and Prognosis of BRAF Mutated Colorectal Cancer Patients. *Frontiers in Oncology* 10, p. 10. doi: 10.3389/fonc.2020.563407

Guo, G., Gong, K., Wohlfeld, B., Hatanpaa, K. J., Zhao, D. and Habib, A. A. 2015. Ligand-Independent EGFR Signaling. *Cancer Res* 75(17), pp. 3436-3441. doi: 10.1158/0008-5472.can-15-0989

References

Guraya, S. Y. 2015. Association of type 2 diabetes mellitus and the risk of colorectal cancer: A meta-analysis and systematic review. *World J Gastroenterol* 21(19), pp. 6026-6031. doi: 10.3748/wjg.v21.i19.6026

Guren, T. K. et al. 2017. Cetuximab in treatment of metastatic colorectal cancer: final survival analyses and extended RAS data from the NORDIC-VII study. *Br J Cancer* 116(10), pp. 1271-1278. doi: 10.1038/bjc.2017.93

Gusev, A. et al. 2016. Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet* 48(3), pp. 245-252. doi: 10.1038/ng.3506

Haggar, F. A. and Boushey, R. P. 2009. Colorectal cancer epidemiology: incidence, mortality, survival, and risk factors. *Clin Colon Rectal Surg* 22(4), pp. 191-197. doi: 10.1055/s-0029-1242458

Hallin, J. et al. 2020. The KRAS(G12C) Inhibitor MRTX849 Provides Insight toward Therapeutic Susceptibility of KRAS-Mutant Cancers in Mouse Models and Patients. *Cancer Discov* 10(1), pp. 54-71. doi: 10.1158/2159-8290.Cd-19-1167

Han, P. et al. 2016. CDH1 rs9929218 variant at 16q22.1 contributes to colorectal cancer susceptibility. *Oncotarget* 7(30), pp. 47278-47286. doi: 10.18632/oncotarget.9758

Hanahan, D. 2022. Hallmarks of Cancer: New Dimensions. *Cancer Discovery* 12(1), pp. 31-46. doi: 10.1158/2159-8290.CD-21-1059

Haydon, A. M. M., Macinnis, R. J., English, D. R. and Giles, G. G. 2006. Effect of physical activity and body size on survival after diagnosis with colorectal cancer. *Gut* 55(1), pp. 62-67. doi: 10.1136/gut.2005.068189

References

He, Y. et al. 2019. Effects of common genetic variants associated with colorectal cancer risk on survival outcomes after diagnosis: A large population-based cohort study. *International Journal of Cancer* 145(9), pp. 2427-2432. doi: 10.1002/ijc.32550

He, Y. Z. et al. 2021. Colorectal cancer risk variants rs10161980 and rs7495132 are associated with cancer survival outcome by a recessive mode of inheritance. *International Journal of Cancer* 148(11), pp. 2774-2778. doi: 10.1002/ijc.33465

Henriksen, L., Grandal, M. V., Knudsen, S. L., van Deurs, B. and Grøvdal, L. M. 2013. Internalization mechanisms of the epidermal growth factor receptor after activation with different ligands. *PLoS One* 8(3), p. e58148. doi: 10.1371/journal.pone.0058148

Hoevenaar, W. H. M. et al. 2020. Degree and site of chromosomal instability define its oncogenic potential. *Nat Commun* 11(1), p. 1501. doi: 10.1038/s41467-020-15279-9

Hormozdiari, F. et al. 2016. Colocalization of GWAS and eQTL Signals Detects Target Genes. *Am J Hum Genet* 99(6), pp. 1245-1260. doi: 10.1016/j.ajhg.2016.10.003

Howie, B., Marchini, J. and Stephens, M. 2011. Genotype Imputation with Thousands of Genomes. *G3: Genes|Genomes|Genetics* 1(6), p. 457. doi: 10.1534/g3.111.001198

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 Genet* 5(6), p. e1000529. doi: 10.1371/journal.pgen.1000529

Hu, Y. et al. 2019. A statistical framework for cross-tissue transcriptome-wide association analysis. *Nature Genetics* 51(3), pp. 568-576. doi: 10.1038/s41588-019-0345-7

References

Huang, Y., Zhao, M., Xu, H., Wang, K., Fu, Z., Jiang, Y. and Yao, Z. 2014. RASAL2 down-regulation in ovarian cancer promotes epithelial-mesenchymal transition and metastasis. *Oncotarget* 5(16), pp. 6734-6745. doi: 10.18632/oncotarget.2244

Hunt, R., Sauna, Z. E., Ambudkar, S. V., Gottesman, M. M. and Kimchi-Sarfaty, C. 2009. Silent (synonymous) SNPs: should we care about them? *Methods Mol Biol* 578, pp. 23-39. doi: 10.1007/978-1-60327-411-1_2

Iacopetta, B. 2002. Are there two sides to colorectal cancer? *International Journal of Cancer* 101(5), pp. 403-408. doi: https://doi.org/10.1002/ijc.10635

Jasperson, K. W., Tuohy, T. M., Neklason, D. W. and Burt, R. W. 2010. Hereditary and familial colon cancer. *Gastroenterology* 138(6), pp. 2044-2058. doi: 10.1053/j.gastro.2010.01.054

Jia, Z., Liu, W., Gong, L. and Xiao, Z. 2017. Downregulation of RASAL2 promotes the proliferation, epithelial-mesenchymal transition and metastasis of colorectal cancer cells. *Oncol Lett* 13(3), pp. 1379-1385. doi: 10.3892/ol.2017.5581

Joachim, C., Macni, J., Drame, M., Pomier, A., Escarmant, P., Veronique-Baudin, J. and Vinh-Hung, V. 2019. Overall survival of colorectal cancer by stage at diagnosis: Data from the Martinique Cancer Registry. *Medicine (Baltimore)* 98(35), p. e16941. doi: 10.1097/md.0000000000016941

Jolliffe, I. T. and Cadima, J. 2016. Principal component analysis: a review and recent developments. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences* 374(2065), pp. 20150202-20150202. doi: 10.1098/rsta.2015.0202

Kalady, M. F. et al. 2012. BRAF Mutations in Colorectal Cancer Are Associated With Distinct Clinical Characteristics and Worse Prognosis. *Diseases of the Colon & Rectum* 55(2), pp. 128-133. doi: 10.1097/DCR.0b013e31823c08b3

References

Karapetis, C. S. et al. 2008a. K-ras Mutations and Benefit from Cetuximab in Advanced Colorectal Cancer. *New England Journal of Medicine* 359(17), pp. 1757-1765. doi: 10.1056/NEJMoa0804385

Karapetis, C. S. et al. 2008b. K-ras mutations and benefit from cetuximab in advanced colorectal cancer. *N Engl J Med* 359(17), pp. 1757-1765. doi: 10.1056/NEJMoa0804385

Kasprzak, A. and Adamek, A. 2019. Insulin-Like Growth Factor 2 (IGF2) Signaling in Colorectal Cancer-From Basic Research to Potential Clinical Applications. *Int J Mol Sci* 20(19), doi: 10.3390/ijms20194915

Kassambara, A., Kosinski, M. and Biecek, P. 2021. survminer: Drawing Survival Curves using 'ggplot2'.

Kastrinos, F. and Syngal, S. 2011. Inherited colorectal cancer syndromes. *Cancer J* 17(6), pp. 405-415. doi: 10.1097/PPO.0b013e318237e408

Kato, S. et al. 2007. PIK3CA mutation is predictive of poor survival in patients with colorectal cancer. *International Journal of Cancer* 121(8), pp. 1771-1778. doi: 10.1002/ijc.22890

Kattan, M. W. et al. 2016. American Joint Committee on Cancer acceptance criteria for inclusion of risk models for individualized prognosis in the practice of precision medicine. *CA Cancer J Clin* 66(5), pp. 370-374. doi: 10.3322/caac.21339

Kawakami, H., Zaanan, A. and Sinicrope, F. A. 2015. Microsatellite instability testing and its role in the management of colorectal cancer. *Curr Treat Options Oncol* 16(7), p. 30. doi: 10.1007/s11864-015-0348-2

References

Keung, E. Z. and Gershenwald, J. E. 2018. The eighth edition American Joint Committee on Cancer (AJCC) melanoma staging system: implications for melanoma treatment and care. *Expert Rev Anticancer Ther* 18(8), pp. 775-784. doi: 10.1080/14737140.2018.1489246

Khattak, M. A., Martin, H., Davidson, A. and Phillips, M. 2015. Role of first-line anti-epidermal growth factor receptor therapy compared with anti-vascular endothelial growth factor therapy in advanced colorectal cancer: a meta-analysis of randomized clinical trials. *Clin Colorectal Cancer* 14(2), pp. 81-90. doi: 10.1016/j.clcc.2014.12.011

Kiang, K. M. Y., Zhang, P. D., Li, N., Zhu, Z. Y., Jin, L. and Leung, G. K. K. 2020. Loss of cytoskeleton protein ADD3 promotes tumor growth and angiogenesis in glioblastoma multiforme. *Cancer Letters* 474, pp. 118-126. doi: 10.1016/j.canlet.2020.01.007

Kiezun, A. et al. 2012. Exome sequencing and the genetic basis of complex traits. *Nature genetics* 44(6), pp. 623-630. doi: 10.1038/ng.2303

Kim, C. H., Huh, J. W., Kim, H. R. and Kim, Y. J. 2017. CpG island methylator phenotype is an independent predictor of survival after curative resection for colorectal cancer: A prospective cohort study. *J Gastroenterol Hepatol* 32(8), pp. 1469-1474. doi: 10.1111/jgh.13734

Kim, J. G. et al. 2008. Vascular endothelial growth factor gene polymorphisms associated with prognosis for patients with colorectal cancer. *Clin Cancer Res* 14(1), pp. 62-66. doi: 10.1158/1078-0432.Ccr-07-1537

Knudson, A. G. 1996. Hereditary cancer: two hits revisited. *J Cancer Res Clin Oncol* 122(3), pp. 135-140. doi: 10.1007/bf01366952

Kraft, P., Zeggini, E. and Ioannidis, J. P. 2009. Replication in genome-wide association studies. *Stat Sci* 24(4), pp. 561-573. doi: 10.1214/09-sts290

References

Kudryavtseva, A. V. et al. 2016. Important molecular genetic markers of colorectal cancer. *Oncotarget* 7(33), pp. 53959-53983. doi: 10.18632/oncotarget.9796

Kuipers, E. J. et al. 2015. Colorectal cancer. *Nat Rev Dis Primers* 1, p. 15065. doi: 10.1038/nrdp.2015.65

Kumuthini, J. et al. 2022. The clinical utility of polygenic risk scores in genomic medicine practices: a systematic review. *Human Genetics* 141(11), pp. 1697-1704. doi: 10.1007/s00439-022-02452-x

Kwong, L. N. and Dove, W. F. 2009. APC and its modifiers in colon cancer. *Adv Exp Med Biol* 656, pp. 85-106. doi: 10.1007/978-1-4419-1145-2_8

Köhne, C. H. et al. 2002. Clinical determinants of survival in patients with 5-fluorouracil-based treatment for metastatic colorectal cancer: results of a multivariate analysis of 3825 patients. *Ann Oncol* 13(2), pp. 308-317. doi: 10.1093/annonc/mdf034

Labadie, J. D. et al. 2022. Genome-wide association study identifies tumor anatomical site-specific risk variants for colorectal cancer survival. *Scientific Reports* 12(1), p. 127. doi: 10.1038/s41598-021-03945-x

Lao, V. V. and Grady, W. M. 2011. Epigenetics and colorectal cancer. *Nat Rev Gastroenterol Hepatol* 8(12), pp. 686-700. doi: 10.1038/nrgastro.2011.173

Lechuga, S., Amin, P. H., Wolen, A. R. and Ivanov, A. I. 2019. Adducins inhibit lung cancer cell migration through mechanisms involving regulation of cell-matrix adhesion and cadherin-11 expression. *Biochimica Et Biophysica Acta-Molecular Cell Research* 1866(3), pp. 395-408. doi: 10.1016/j.bbamcr.2018.10.001

References

Lee, J. C., Wang, S. T., Chow, N. H. and Yang, H. B. 2002. Investigation of the prognostic value of coexpressed erbB family members for the survival of colorectal cancer patients after curative surgery. *Eur J Cancer* 38(8), pp. 1065-1071. doi: 10.1016/s0959-8049(02)00004-7

Leeuw, C., Sey, N., Posthuma, D. and Won, H. 2020. *A response to Yurko et al: H-MAGMA, inheriting a shaky statistical foundation, yields excess false positives*.

Leitch, E. F., Chakrabarti, M., Crozier, J. E., McKee, R. F., Anderson, J. H., Horgan, P. G. and McMillan, D. C. 2007. Comparison of the prognostic value of selected markers of the systemic inflammatory response in patients with colorectal cancer. *Br J Cancer* 97(9), pp. 1266-1270. doi: 10.1038/sj.bjc.6604027

Leslie, A., Carey, F. A., Pratt, N. R. and Steele, R. J. 2002. The colorectal adenoma-carcinoma sequence. *Br J Surg* 89(7), pp. 845-860. doi: 10.1046/j.1365-2168.2002.02120.x

Lewontin, R. C. 1964. The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. *Genetics* 49(1), pp. 49-67. doi: 10.1093/genetics/49.1.49

Li, B. and Ritchie, M. D. 2021. From GWAS to Gene: Transcriptome-Wide Association Studies and Other Methods to Functionally Understand GWAS Discoveries. *Frontiers in Genetics* 12,

Li, L. et al. 2017. Investigation of correlation between mutational status in key EGFR signaling genes and prognosis of stage II colorectal cancer. *Future Oncology* 13(17), pp. 1473-1492. doi: 10.2217/fon-2017-0040

References

Li, N. and Li, S. 2014. RASAL2 promotes lung cancer metastasis through epithelial-mesenchymal transition. *Biochem Biophys Res Commun* 455(3-4), pp. 358-362. doi: 10.1016/j.bbrc.2014.11.020

Li, N., Lorenzi, F., Kalakouti, E., Normatova, M., Babaei-Jadidi, R., Tomlinson, I. and Nateri, A. S. 2015a. FBXW7-mutated colorectal cancer cells exhibit aberrant expression of phosphorylated-p53 at Serine-15. *Oncotarget* 6(11), pp. 9240-9256. doi: 10.18632/oncotarget.3284

Li, R., Liang, M., Liang, X., Yang, L., Su, M. and Lai, K. P. 2020. Chemotherapeutic Effectiveness of Combining Cetuximab for Metastatic Colorectal Cancer Treatment: A System Review and Meta-Analysis. *Front Oncol* 10, p. 868. doi: 10.3389/fonc.2020.00868

Li, W. B. et al. 2015b. Colorectal carcinomas with KRAS codon 12 mutation are associated with more advanced tumor stages. *Bmc Cancer* 15, p. 9. doi: 10.1186/s12885-015-1345-3

Li, Y., Willer, C., Sanna, S. and Abecasis, G. 2009. Genotype Imputation. *Annual Review of Genomics and Human Genetics* 10(1), pp. 387-406. doi: 10.1146/annurev.genom.9.081307.164242

Liang, P. S., Chen, T. Y. and Giovannucci, E. 2009. Cigarette smoking and colorectal cancer incidence and mortality: systematic review and meta-analysis. *Int J Cancer* 124(10), pp. 2406-2415. doi: 10.1002/ijc.24191

Liebl, M. C. and Hofmann, T. G. 2021. The Role of p53 Signaling in Colorectal Cancer. *Cancers (Basel)* 13(9), doi: 10.3390/cancers13092125

Liew, C.-C., Ma, J., Tang, H.-C., Zheng, R. and Dempsey, A. A. 2006. The peripheral blood transcriptome dynamically reflects system wide biology: a potential diagnostic

tool. *Journal of Laboratory and Clinical Medicine* 147(3), pp. 126-132. doi: https://doi.org/10.1016/j.lab.2005.10.005

Limsui, D. et al. 2012. Postmenopausal hormone therapy and colorectal cancer risk by molecularly defined subtypes among older women. *Gut* 61(9), pp. 1299-1305. doi: 10.1136/gutjnl-2011-300719

Lito, P., Solomon, M., Li, L.-S., Hansen, R. and Rosen, N. 2016. Allele-specific inhibitors inactivate mutant KRAS G12C by a trapping mechanism. *Science* 351(6273), pp. 604-608. doi: 10.1126/science.aad6204

Liu, J. et al. 2021. Genome-wide association study followed by trans-ancestry meta-analysis identify 17 new risk loci for schizophrenia. *BMC Medicine* 19(1), p. 177. doi: 10.1186/s12916-021-02039-9

Liu, Y. et al. 2014. Association between statin use and colorectal cancer risk: a meta-analysis of 42 studies. *Cancer Causes Control* 25(2), pp. 237-249. doi: 10.1007/s10552-013-0326-6

Lochhead, P. et al. 2013. Microsatellite instability and BRAF mutation testing in colorectal cancer prognostication. *J Natl Cancer Inst* 105(15), pp. 1151-1156. doi: 10.1093/jnci/djt173

Lu, Y. et al. 2019. Large-Scale Genome-Wide Association Study of East Asians Identifies Loci Associated With Risk for Colorectal Cancer. *Gastroenterology* 156(5), pp. 1455-1466. doi: 10.1053/j.gastro.2018.11.066

Luo, C. and Shen, J. Y. 2017. Adducin in tumorigenesis and metastasis. *Oncotarget* 8(29), pp. 48453-48459. doi: 10.18632/oncotarget.17173

References

Luo, Y. et al. 2013. RET is a potential tumor suppressor gene in colorectal cancer. *Oncogene* 32(16), pp. 2037-2047. doi: 10.1038/onc.2012.225

Lv, W., Zhang, G. Q., Jiao, A., Zhao, B. C., Shi, Y., Chen, B. M. and Zhang, J. L. 2017. Chemotherapy Plus Cetuximab versus Chemotherapy Alone for Patients with KRAS Wild Type Unresectable Liver-Confined Metastases Colorectal Cancer: An Updated Meta-Analysis of RCTs. *Gastroenterol Res Pract* 2017, p. 8464905. doi: 10.1155/2017/8464905

Lynch, H. T. and de la Chapelle, A. 2003. Hereditary colorectal cancer. *N Engl J Med* 348(10), pp. 919-932. doi: 10.1056/NEJMra012242

Lynch, H. T., Lynch, P. M., Lanspa, S. J., Snyder, C. L., Lynch, J. F. and Boland, C. R. 2009. Review of the Lynch syndrome: history, molecular genetics, screening, differential diagnosis, and medicolegal ramifications. *Clin Genet* 76(1), pp. 1-18. doi: 10.1111/j.1399-0004.2009.01230.x

Majek, O. et al. 2013. Sex differences in colorectal cancer survival: population-based analysis of 164,996 colorectal cancer patients in Germany. *PLoS One* 8(7), p. e68077. doi: 10.1371/journal.pone.0068077

Manolio, T. A. et al. 2009. Finding the missing heritability of complex diseases. *Nature* 461(7265), pp. 747-753. doi: 10.1038/nature08494

Marchini, J., Cardon, L. R., Phillips, M. S. and Donnelly, P. 2004. The effects of human population structure on large genetic association studies. *Nature Genetics* 36(5), pp. 512-517. doi: 10.1038/ng1337

Marcuello, E., Altes, A., del Rio, E., Cesar, A., Menoyo, A. and Baiget, M. 2004. Single nucleotide polymorphism in the 5' tandem repeat sequences of thymidylate synthase

gene predicts for response to fluorouracil-based chemotherapy in advanced colorectal cancer patients. *Int J Cancer* 112(5), pp. 733-737. doi: 10.1002/ijc.20487

Marees, A. T., de Kluiver, H., Stringer, S., Vorspan, F., Curis, E., Marie-Claire, C. and Derks, E. M. 2018a. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *Int J Methods Psychiatr Res* 27(2), p. e1608. doi: 10.1002/mpr.1608

Marees, A. T., de Kluiver, H., Stringer, S., Vorspan, F., Curis, E., Marie-Claire, C. and Derks, E. M. 2018b. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International Journal of Methods in Psychiatric Research* 27(2), p. e1608. doi: 10.1002/mpr.1608

Matsuo, H. et al. 2016. Genome-wide association study of clinically defined gout identifies multiple risk loci and its association with clinical subtypes. *Ann Rheum Dis* 75(4), pp. 652-659. doi: 10.1136/annrheumdis-2014-206191

Maughan, T. S. et al. 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), pp. 2103-2114. doi: 10.1016/s0140-6736(11)60613-2

McCarthy, S. et al. 2016. A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics* 48(10), pp. 1279-1283. doi: 10.1038/ng.3643

McKay, A. et al. 2014. Does young age influence the prognosis of colorectal cancer: a population-based analysis. *World Journal of Surgical Oncology* 12(1), p. 370. doi: 10.1186/1477-7819-12-370

McKinney, W. and others eds. 2010. *Data structures for statistical computing in python*. Austin, TX.

References

McLaughlin, S. K. et al. 2013. The RasGAP gene, RASAL2, is a tumor and metastasis suppressor. *Cancer Cell* 24(3), pp. 365-378. doi: 10.1016/j.ccr.2013.08.004

Meguid, R. A., Slidell, M. B., Chang, D. C. and Ahuja, N. 2007. Is there a difference in survival between right- versus left-sided colon cancers? *Annals of Surgical Oncology* 14(2), pp. 96-96.

Mehta, D. et al. 2013. Impact of common regulatory single-nucleotide variants on gene expression profiles in whole blood. *European Journal of Human Genetics* 21(1), pp. 48-54. doi: 10.1038/ejhg.2012.106

Mendelsohn, J., Prewett, M., Rockwell, P. and Goldstein, N. I. 2015. CCR 20th anniversary commentary: a chimeric antibody, C225, inhibits EGFR activation and tumor growth.*Clin Cancer Res*. Vol. 21. United States: ©2015 American Association for Cancer Research., pp. 227-229.

Meng, M., Zhong, K., Jiang, T., Liu, Z., Kwan, H. Y. and Su, T. 2021. The current understanding on the impact of KRAS on colorectal cancer. *Biomedicine & Pharmacotherapy* 140, p. 111717. doi: https://doi.org/10.1016/j.biopha.2021.111717

Min, J. et al. 2010. An oncogene-tumor suppressor cascade drives metastatic prostate cancer by coordinately activating Ras and nuclear factor-kappaB. *Nature medicine* 16(3), pp. 286-294. doi: 10.1038/nm.2100

Missiaglia, E. et al. 2014. Distal and proximal colon cancers differ in terms of molecular, pathological, and clinical features. *Annals of Oncology* 25(10), pp. 1995-2001. doi: 10.1093/annonc/mdu275

References

Mizuno, A. and Okada, Y. 2019. Biological characterization of expression quantitative trait loci (eQTLs) showing tissue-specific opposite directional effects. *European Journal of Human Genetics* 27(11), pp. 1745-1756. doi: 10.1038/s41431-019-0468-4

Morris, E. J. et al. 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), p. e0117816. doi: 10.1371/journal.pone.0117816

Muller, Chesner, Egan, Rowlands, Collard, Swarbrick and Newman. 1989. Significance of venous and lymphatic invasion in malignant polyps of the colon and rectum. *Gut* 30(10), p. 1385. doi: 10.1136/gut.30.10.1385

Muzny, D. M. et al. 2012. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 487(7407), pp. 330-337. doi: 10.1038/nature11252

Müller, D. and Győrffy, B. 2022. DNA methylation-based diagnostic, prognostic, and predictive biomarkers in colorectal cancer. *Biochim Biophys Acta Rev Cancer* 1877(3), p. 188722. doi: 10.1016/j.bbcan.2022.188722

Negrini, S., Gorgoulis, V. G. and Halazonetis, T. D. 2010. Genomic instability--an evolving hallmark of cancer. *Nat Rev Mol Cell Biol* 11(3), pp. 220-228. doi: 10.1038/nrm2858

Newcomb, P. A. et al. 2007. Colon cancer family registry: An international resource for studies of the genetic epidemiology of colon cancer. *Cancer Epidemiology Biomarkers & Prevention* 16(11), pp. 2331-2343. doi: 10.1158/1055-9965.epi-07-0648

Nica, A. C. and Dermitzakis, E. T. 2013. Expression quantitative trait loci: present and future. *Philos Trans R Soc Lond B Biol Sci* 368(1620), p. 20120362. doi: 10.1098/rstb.2012.0362

References

null, n. et al. 2020. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* 369(6509), pp. 1318-1330. doi: 10.1126/science.aaz1776

Oda, K., Matsuoka, Y., Funahashi, A. and Kitano, H. 2005. A comprehensive pathway map of epidermal growth factor receptor signaling. *Mol Syst Biol* 1, p. 2005.0010. doi: 10.1038/msb4100014

Ogino, S. et al. 2006. CpG island methylator phenotype (CIMP) of colorectal cancer is best characterised by quantitative DNA methylation analysis and prospective cohort studies. *Gut* 55(7), pp. 1000-1006. doi: 10.1136/gut.2005.082933

Ogino, S. et al. 2009. CpG island methylator phenotype, microsatellite instability, BRAF mutation and clinical outcome in colon cancer. *Gut* 58(1), pp. 90-96. doi: 10.1136/gut.2008.155473

Ogunbiyi, O. A. et al. 1998. Confirmation that chromosome 18q allelic loss in colon cancer is a prognostic indicator. *Journal of Clinical Oncology* 16(2), pp. 427-433. doi: 10.1200/JCO.1998.16.2.427

Ogura, T. et al. 2014. Clinicopathological characteristics and prognostic impact of colorectal cancers with NRAS mutations. *Oncology Reports* 32(1), pp. 50-56. doi: 10.3892/or.2014.3165

Oliver, P., Zachary, G., Eske, D., Naomi, R. W., Alexander, G. and Ammar, A.-C. 2022. Polygenic Prediction of Molecular Traits using Large-Scale Meta-analysis Summary Statistics. *bioRxiv*, p. 2022.2011.2023.517213. doi: 10.1101/2022.11.23.517213

Olsen, S. N. et al. 2017. Loss of RasGAP Tumor Suppressors Underlies the Aggressive Nature of Luminal B Breast Cancers. *Cancer Discov* 7(2), pp. 202-217. doi: 10.1158/2159-8290.cd-16-0520

References

Orsetti, B. et al. 2014. Impact of chromosomal instability on colorectal cancer progression and outcome. *BMC Cancer* 14(1), p. 121. doi: 10.1186/1471-2407-14-121

Owzar, K., Li, Z., Cox, N. and Jung, S.-H. 2012. Power and Sample Size Calculations for SNP Association Studies With Censored Time-to-Event Outcomes. *Genetic Epidemiology* 36, pp. 538-548.

Pan, Y. et al. 2018. RASAL2 promotes tumor progression through LATS2/YAP1 axis of hippo signaling pathway in colorectal cancer. *Molecular Cancer* 17(1), p. 102. doi: 10.1186/s12943-018-0853-6

Parikshak, N. N. and Geschwind, D. H. 2013. Chapter 84 - Neuroscience and the Genomic Revolution: An Overview. In: Ginsburg, G.S. and Willard, H.F. eds. *Genomic and Personalized Medicine (Second Edition)*. Academic Press, pp. 1018-1027.

Pectasides, E. and Bass, A. J. 2015. ERBB2 emerges as a new target for colorectal cancer. *Cancer Discov* 5(8), pp. 799-801. doi: 10.1158/2159-8290.cd-15-0730

Peng, C. X., Guo, Z. J., Wu, X. Y. and Zhang, X. L. 2015. A polymorphism at the microRNA binding site in the 3 ' untranslated region of RYR3 is associated with outcome in hepatocellular carcinoma. *Oncotargets and Therapy* 8, pp. 2075-2079. doi: 10.2147/ott.s85856

Penney, M. E., Parfrey, P. S., Savas, S. and Yilmaz, Y. E. 2019. A genome-wide association study identifies single nucleotide polymorphisms associated with time-to-metastasis in colorectal cancer. *BMC Cancer* 19(1), p. 133. doi: 10.1186/s12885-019-5346-5

Perneger, T. V. 1998. What's wrong with Bonferroni adjustments. *British Medical Journal* 316(7139), pp. 1236-1238. doi: 10.1136/bmj.316.7139.1236

References

Phipps, A. I. et al. 2013. Colon and Rectal Cancer Survival by Tumor Location and Microsatellite Instability: The Colon Cancer Family Registry. *Diseases of the Colon & Rectum* 56(8), pp. 937-944. doi: 10.1097/DCR.0b013e31828f9a57

Phipps, A. I. et al. 2012. Association between colorectal cancer susceptibility loci and survival time after diagnosis with colorectal cancer. *Gastroenterology* 143(1), pp. 51-54.e54. doi: 10.1053/j.gastro.2012.04.052

Phipps, A. I. et al. 2016. Common genetic variation and survival after colorectal cancer diagnosis: a genome-wide analysis. *Carcinogenesis* 37(1), pp. 87-95. doi: 10.1093/carcin/bgv161

Pino, M. S. and Chung, D. C. 2010. The chromosomal instability pathway in colon cancer. *Gastroenterology* 138(6), pp. 2059-2072. doi: 10.1053/j.gastro.2009.12.065

Platz, E. A., Willett, W. C., Colditz, G. A., Rimm, E. B., Spiegelman, D. and Giovannucci, E. 2000. Proportion of colon cancer risk that might be preventable in a cohort of middle-aged US men. *Cancer Causes Control* 11(7), pp. 579-588. doi: 10.1023/a:1008999232442

Popat, S., Hubner, R. and Houlston, R. S. 2005. Systematic review of microsatellite instability and colorectal cancer prognosis. *J Clin Oncol* 23(3), pp. 609-618. doi: 10.1200/jco.2005.01.086

Porru, M., Pompili, L., Caruso, C., Biroccio, A. and Leonetti, C. 2018. Targeting KRAS in metastatic colorectal cancer: current strategies and emerging opportunities. *Journal of experimental & clinical cancer research : CR* 37(1), pp. 57-57. doi: 10.1186/s13046-018-0719-1

References

Prorok, P. C. et al. 2000. Design of the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial. *Controlled Clinical Trials* 21(6), pp. 273S-309S. doi: 10.1016/s0197-2456(00)00098-2

Purcell, S. et al. 2007. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics* 81(3), pp. 559-575. doi: https://doi.org/10.1086/519795

Rawla, P., Sunkara, T. and Barsouk, A. 2019. Epidemiology of colorectal cancer: incidence, mortality, survival, and risk factors. *Prz Gastroenterol* 14(2), pp. 89-103. doi: 10.5114/pg.2018.81072

Reeves, G. K., Pirie, K., Beral, V., Green, J., Spencer, E. and Bull, D. 2007. Cancer incidence and mortality in relation to body mass index in the Million Women Study: cohort study. *BMJ* 335(7630), p. 1134. doi: 10.1136/bmj.39367.495995.AE

Ren, W. G., Sun, Z. Q., Zeng, Q. L., Han, S., Zhang, Q. L. and Jiang, L. B. 2016. Aberrant Expression of CUL4A Is Associated with IL-6/STAT3 Activation in Colorectal Cancer Progression. *Archives of Medical Research* 47(3), pp. 214-222. doi: 10.1016/j.arcmed.2016.07.001

Revelle, W. 2021. *psych: Procedures for Psychological, Psychometric, and Personality Research*. Available at: https://CRAN.R-project.org/package=psych [Accessed.

Riboli, E. and Kaaks, R. 1997. The EPIC Project: rationale and study design. European Prospective Investigation into Cancer and Nutrition. *Int J Epidemiol* 26 Suppl 1, pp. S6-14. doi: 10.1093/ije/26.suppl_1.s6

Richard, G. F., Kerrest, A. and Dujon, B. 2008. Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. *Microbiol Mol Biol Rev* 72(4), pp. 686-727. doi: 10.1128/mmbr.00011-08

References

Richman, S. D. et al. 2009. KRAS and BRAF Mutations in Advanced Colorectal Cancer Are Associated With Poor Prognosis but Do Not Preclude Benefit From Oxaliplatin or Irinotecan: Results From the MRC FOCUS Trial. *Journal of Clinical Oncology* 27(35), pp. 5931-5937. doi: 10.1200/jco.2009.22.4295

Riihimäki, M., Hemminki, A., Sundquist, J. and Hemminki, K. 2016. Patterns of metastasis in colon and rectal cancer. *Sci Rep* 6, p. 29765. doi: 10.1038/srep29765

Rimm, E. B., Giovannucci, E. L., Willett, W. C., Colditz, G. A., Ascherio, A., Rosner, B. and Stampfer, M. J. 1991. PROSPECTIVE-STUDY OF ALCOHOL-CONSUMPTION AND RISK OF CORONARY-DISEASE IN MEN. *Lancet* 338(8765), pp. 464-468. doi: 10.1016/0140-6736(91)90542-w

Rizvi, A. A. et al. 2019. gwasurvivr: an R package for genome-wide survival analysis. *Bioinformatics* 35(11), pp. 1968-1970. doi: 10.1093/bioinformatics/bty920

Rosty, C. et al. 2013. Colorectal carcinomas with KRAS mutation are associated with distinctive morphological and molecular features. *Modern Pathology* 26(6), pp. 825-834. doi: 10.1038/modpathol.2012.240

R_Core_Team. 2018. R: A Language and Environment for Statistical Computing.

Sabari, J. K. et al. 2021. KRYSTAL-2: A phase I/II trial of adagrasib (MRTX849) in combination with TNO155 in patients with advanced solid tumors with KRAS G12C mutation. *Journal of Clinical Oncology* 39(3_suppl), pp. TPS146-TPS146. doi: 10.1200/JCO.2021.39.3_suppl.TPS146

Saif, M. W., Alexander, D. and Wicox, C. M. 2005. Serum Alkaline Phosphatase Level as a Prognostic Tool in Colorectal Cancer: A Study of 105 patients. *J Appl Res* 5(1), pp. 88-95.

References

Salvatore, L. et al. 2019. PTEN in Colorectal Cancer: Shedding Light on Its Role as Predictor and Target. *Cancers (Basel)* 11(11), doi: 10.3390/cancers11111765

Sansregret, L., Vanhaesebroeck, B. and Swanton, C. 2018. Determinants and clinical implications of chromosomal instability in cancer. *Nat Rev Clin Oncol* 15(3), pp. 139-150. doi: 10.1038/nrclinonc.2017.198

Sanz-Pamplona, R. et al. 2011. Gene Expression Differences between Colon and Rectum Tumors. *Clinical Cancer Research* 17(23), pp. 7303-7312. doi: 10.1158/1078-0432.ccr-11-1570

Sassano, M., Mariani, M., Quaranta, G., Pastorino, R. and Boccia, S. 2022. Polygenic risk prediction models for colorectal cancer: a systematic review. *BMC Cancer* 22(1), p. 65. doi: 10.1186/s12885-021-09143-2

Schirripa, M. et al. 2019. Class 1, 2, and 3 BRAF-Mutated Metastatic Colorectal Cancer: A Detailed Clinical, Pathologic, and Molecular Characterization. *Clin Cancer Res* 25(13), pp. 3954-3961. doi: 10.1158/1078-0432.ccr-19-0311

Schirripa, M. et al. 2015. Role of NRAS mutations as prognostic and predictive markers in metastatic colorectal cancer. *International Journal of Cancer* 136(1), pp. 83-90. doi: 10.1002/ijc.28955

Schmuck, R. et al. 2020. Gender comparison of clinical, histopathological, therapeutic and outcome factors in 185,967 colon cancer patients. *Langenbecks Arch Surg* 405(1), pp. 71-80. doi: 10.1007/s00423-019-01850-6

Setu, T. J. and Basak, T. 2021. An Introduction to Basic Statistical Models in Genetics. *Open Journal of Statistics* 11(6), pp. 1017-1025. doi: 10.4236/ojs.2021.116060 .

References

Sey, N. Y. A. et al. 2020. A computational tool (H-MAGMA) for improved prediction of brain-disorder risk genes by incorporating brain chromatin interaction profiles. *Nature Neuroscience* 23(4), pp. 583-593. doi: 10.1038/s41593-020-0603-0

Sham, P. C. and Purcell, S. M. 2014. Statistical power and significance testing in large-scale genetic studies. *Nature Reviews Genetics* 15(5), pp. 335-346. doi: 10.1038/nrg3706

Shao, H. J., Li, Q., Shi, T., Zhang, G. Z. and Shao, F. 2019. LINC00707 promotes cell proliferation and invasion of colorectal cancer via miR-206/FMNL2 axis. *Eur Rev Med Pharmacol Sci* 23(9), pp. 3749-3759. doi: 10.26355/eurrev_201905_17801

Sinicrope, F. A. 2010. DNA mismatch repair and adjuvant chemotherapy in sporadic colon cancer.*Nat Rev Clin Oncol*. Vol. 7. England, pp. 174-177.

Sinicrope, F. A. and Sargent, D. J. 2012. Molecular pathways: microsatellite instability in colorectal cancer: prognostic, predictive, and therapeutic implications. *Clin Cancer Res* 18(6), pp. 1506-1512. doi: 10.1158/1078-0432.ccr-11-1469

Skoulidis, F. et al. 2021. Sotorasib for Lung Cancers with KRAS p.G12C Mutation. *New England Journal of Medicine* 384(25), pp. 2371-2381. doi: 10.1056/NEJMoa2103695

Slatkin, M. 2008. Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics* 9(6), pp. 477-485. doi: 10.1038/nrg2361

Slattery, M. L., Caan, B. J., Benson, J. and Murtaugh, M. 2003. Energy balance and rectal cancer: An evaluation of energy intake, energy expenditure, and body mass index. *Nutrition and Cancer-an International Journal* 46(2), pp. 166-171. doi: 10.1207/s15327914nc4602_09

References

Slattery, M. L., Potter, J., Caan, B., Edwards, S., Coates, A., Ma, K. N. and Berry, T. D. 1997. Energy balance and colon cancer - Beyond physical activity. *Cancer Research* 57(1), pp. 75-80.

Smith, C. G. et al. 2013. Somatic profiling of the epidermal growth factor receptor pathway in tumors from patients with advanced colorectal cancer treated with chemotherapy +/- cetuximab. *Clin Cancer Res* 19(15), pp. 4104-4113. doi: 10.1158/1078-0432.Ccr-12-2581

Smith, C. G. et al. 2015. Analyses of 7,635 Patients with Colorectal Cancer Using Independent Training and Validation Cohorts Show That rs9929218 in CDH1 Is a Prognostic Marker of Survival. *Clin Cancer Res* 21(15), pp. 3453-3461. doi: 10.1158/1078-0432.Ccr-14-3136

Song, L. and Li, Y. 2015. SEPT9: A Specific Circulating Biomarker for Colorectal Cancer. *Adv Clin Chem* 72, pp. 171-204. doi: 10.1016/bs.acc.2015.07.004

Song, M., Garrett, W. S. and Chan, A. T. 2015. Nutrients, foods, and colorectal cancer prevention. *Gastroenterology* 148(6), pp. 1244-1260.e1216. doi: 10.1053/j.gastro.2014.12.035

Song, N. et al. 2018. Colorectal cancer susceptibility loci and influence on survival. *Genes Chromosomes & Cancer* 57(12), pp. 630-637. doi: 10.1002/gcc.22674

Srivastava, S. and Wagner, P. D. 2020. The Early Detection Research Network: A National Infrastructure to Support the Discovery, Development, and Validation of Cancer Biomarkers. *Cancer Epidemiology Biomarkers & Prevention* 29(12), pp. 2401-2410. doi: 10.1158/1055-9965.epi-20-0237

References

Steering-Committee. 1989. Final report on the aspirin component of the ongoing Physicians' Health Study. *N Engl J Med* 321(3), pp. 129-135. doi: 10.1056/nejm198907203210301

Stefanska, B. et al. 2014. Genome-wide study of hypomethylated and induced genes in patients with liver cancer unravels novel anticancer targets. *Clin Cancer Res* 20(12), pp. 3118-3132. doi: 10.1158/1078-0432.Ccr-13-0283

Stelzner, S., Hellmich, G., Koch, R. and Ludwig, K. 2005. Factors predicting survival in stage IV colorectal carcinoma patients after palliative treatment: a multivariate analysis. *J Surg Oncol* 89(4), pp. 211-217. doi: 10.1002/jso.20196

Stintzing, S. et al. 2016. FOLFIRI plus cetuximab versus FOLFIRI plus bevacizumab for metastatic colorectal cancer (FIRE-3): a post-hoc analysis of tumour dynamics in the final RAS wild-type subgroup of this randomised open-label phase 3 trial. *Lancet Oncol* 17(10), pp. 1426-1434. doi: 10.1016/s1470-2045(16)30269-8

Storey, J. D., Bass, A. J., Dabney, A. and Robinson, D. 2021. qvalue: Q-value estimation for false discovery rate control.

Strandberg Holka, P., Eriksson, S., Eberhard, J., Bergenfeldt, M., Lindell, G. and Sturesson, C. 2018. Significance of poor performance status after resection of colorectal liver metastases. *World Journal of Surgical Oncology* 16(1), p. 3. doi: 10.1186/s12957-017-1306-1

Strippoli, A. et al. 2020. c-MYC Expression Is a Possible Keystone in the Colorectal Cancer Resistance to EGFR Inhibitors. *Cancers (Basel)* 12(3), doi: 10.3390/cancers12030638

References

Sud, A., Horton, R. H., Hingorani, A. D., Tzoulaki, I., Turnbull, C., Houlston, R. S. and Lucassen, A. 2023. Realistic expectations are key to realising the benefits of polygenic scores. *BMJ* 380, p. e073149. doi: 10.1136/bmj-2022-073149

Sui, X. M., Zhou, H., Zhu, L., Wang, D. Q., Fan, S. M. and Zhao, W. 2017. CUL4A promotes proliferation and metastasis of colorectal cancer cells by regulating H3K4 trimethylation in epithelial-mesenchymal transition. *Oncotargets and Therapy* 10, pp. 735-743. doi: 10.2147/ott.s118897

Sveen, A., Kopetz, S. and Lothe, R. A. 2020. Biomarker-guided therapy for colorectal cancer: strength in complexity. *Nat Rev Clin Oncol* 17(1), pp. 11-32. doi: 10.1038/s41571-019-0241-1

Swets, J. A. 1988. Measuring the accuracy of diagnostic systems. *Science* 240(4857), pp. 1285-1293. doi: 10.1126/science.3287615

Tailor, K., Paul, J., Ghosh, S., Kumari, N. and Kwabi-Addo, B. 2021. RASAL2 suppresses the proliferative and invasive ability of PC3 prostate cancer cells. *Oncotarget* 12(26), pp. 2489-2499. doi: 10.18632/oncotarget.28158

Takatsuno, Y. et al. 2013. The rs6983267 SNP is associated with MYC transcription efficiency, which promotes progression and worsens prognosis of colorectal cancer. *Ann Surg Oncol* 20(4), pp. 1395-1402. doi: 10.1245/s10434-012-2657-z

Takeichi, M. 1991. Cadherin cell adhesion receptors as a morphogenetic regulator. *Science* 251(5000), pp. 1451-1455. doi: 10.1126/science.2006419

Tatijana, Z. and Vesna, B. 2011. Genetics of Type 1 Diabetes. In: David, W. ed. *Type 1 Diabetes*. Rijeka: IntechOpen, p. Ch. 23.

References

Tenesa, A. et al. 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), pp. 631-637. doi: 10.1038/ng.133

Terradas, M. et al. 2023. MBD4-associated neoplasia syndrome: screening of cases with suggestive phenotypes. *European Journal of Human Genetics*, doi: 10.1038/s41431-023-01418-5

Testa, U., Pelosi, E. and Castelli, G. 2018. Colorectal cancer: genetic abnormalities, tumor progression, tumor heterogeneity, clonal evolution and tumor-initiating cells. *Med Sci (Basel)* 6(2), doi: 10.3390/medsci6020031

Theodoratou, E. et al. 2018. Genome-wide scan of the effect of common nsSNPs on colorectal cancer survival outcome. *British Journal of Cancer* 119(8), pp. 988-993. doi: 10.1038/s41416-018-0117-7

Theodoratou, E. et al. 2007. Dietary flavonoids and the risk of colorectal cancer. *Cancer Epidemiology Biomarkers & Prevention* 16(4), pp. 684-693. doi: 10.1158/1055-9965.epi-06-0785

Therneau, T. M. 2022. A Package for Survival Analysis in R.

Torry, D. S. and Cooper, G. M. 1991. Proto-Oncogenes in Development and Cancer. *American Journal of Reproductive Immunology* 25(3), pp. 129-132. doi: https://doi.org/10.1111/j.1600-0897.1991.tb01080.x

Tosti, E. et al. 2022. Loss of MMR and TGFBR2 Increases the Susceptibility to Microbiota-Dependent Inflammation-Associated Colon Cancer. *Cell Mol Gastroenterol Hepatol* 14(3), pp. 693-717. doi: 10.1016/j.jcmgh.2022.05.010

References

Toyota, M., Ahuja, N., Ohe-Toyota, M., Herman, J. G., Baylin, S. B. and Issa, J. P. 1999. CpG island methylator phenotype in colorectal cancer. *Proc Natl Acad Sci U S A* 96(15), pp. 8681-8686. doi: 10.1073/pnas.96.15.8681

Tran, B. et al. 2011a. Impact of BRAF mutation and microsatellite instability on the pattern of metastatic spread and prognosis in metastatic colorectal cancer. *Cancer* 117(20), pp. 4623-4632. doi: https://doi.org/10.1002/cncr.26086

Tran, B. et al. 2011b. Impact of BRAF Mutation and Microsatellite Instability on the Pattern of Metastatic Spread and Prognosis in Metastatic Colorectal Cancer. *Cancer* 117(20), pp. 4623-4632. doi: 10.1002/cncr.26086

Travaglino, A. et al. 2019. Clinicopathological factors associated with BRAF-V600E mutation in colorectal serrated adenomas. *Histopathology* 75(2), pp. 160-173. doi: 10.1111/his.13846

Turner, S. 2018. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *The Journal of Open Source Software*, doi: 10.21105/joss.00731

Turner, S. et al. 2011. Quality control procedures for genome-wide association studies. *Current protocols in human genetics* Chapter 1, pp. Unit1.19-Unit11.19. doi: 10.1002/0471142905.hg0119s68

Tuupanen, S. et al. 2009. The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. *Nat Genet* 41(8), pp. 885-890. doi: 10.1038/ng.406

Uffelmann, E. et al. 2021. Genome-wide association studies. *Nature Reviews Methods Primers* 1(1), p. 59. doi: 10.1038/s43586-021-00056-9

References

Uhlen, M. et al. 2015. Tissue-based map of the human proteome. *Science* 347(6220), p. 10. doi: 10.1126/science.1260419

Uhlen, M. et al. 2017. A pathology atlas of the human cancer transcriptome. *Science* 357(6352), pp. 660-+. doi: 10.1126/science.aan2507

Urmo, V. et al. 2018. Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. *bioRxiv*, p. 447367. doi: 10.1101/447367

Van Cutsem, E. et al. 2015. Fluorouracil, Leucovorin, and Irinotecan Plus Cetuximab Treatment and RAS Mutations in Colorectal Cancer. *Journal of Clinical Oncology* 33(7), pp. 692-700. doi: 10.1200/JCO.2014.59.4812

van Eeghen, E. E., Bakker, S. D., van Bochove, A. and Loffeld, R. J. 2015. Impact of age and comorbidity on survival in colorectal cancer. *J Gastrointest Oncol* 6(6), pp. 605-612. doi: 10.3978/j.issn.2078-6891.2015.070

Van Rossum, G. and Drake, F. L. 2009. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A. and Yang, J. 2017. 10 Years of GWAS Discovery: Biology, Function, and Translation. *The American Journal of Human Genetics* 101(1), pp. 5-22. doi: https://doi.org/10.1016/j.ajhg.2017.06.005

Vogelstein, B. and Kinzler, K. W. 2004. Cancer genes and the pathways they control. *Nat Med* 10(8), pp. 789-799. doi: 10.1038/nm1087

Wainberg, M. et al. 2019. Opportunities and challenges for transcriptome-wide association studies. *Nature Genetics* 51(4), pp. 592-599. doi: 10.1038/s41588-019-0385-z

Walther, A., Houlston, R. and Tomlinson, I. 2008. Association between chromosomal instability and prognosis in colorectal cancer: a meta-analysis. *Gut* 57(7), pp. 941-950. doi: 10.1136/gut.2007.135004

Walther, A., Johnstone, E., Swanton, C., Midgley, R., Tomlinson, I. and Kerr, D. 2009. Genetic prognostic and predictive markers in colorectal cancer. *Nature Reviews Cancer* 9(7), pp. 489-499. doi: 10.1038/nrc2645

Wan, S. et al. 2013. Preoperative platelet count associates with survival and distant metastasis in surgically resected colorectal cancer patients. *J Gastrointest Cancer* 44(3), pp. 293-304. doi: 10.1007/s12029-013-9491-9

Wang, H., Luan, H., Zhan, T., Liu, X., Song, J. and Dai, H. 2020. Long non-coding RNA LINC00707 acts as a competing endogenous RNA to enhance cell proliferation in colorectal cancer. *Exp Ther Med* 19(2), pp. 1439-1447. doi: 10.3892/etm.2019.8350

Wang, W. M. et al. 2015. Cullin1 is a novel prognostic marker and regulates the cell proliferation and metastasis in colorectal cancer. *Journal of Cancer Research and Clinical Oncology* 141(9), pp. 1603-1612. doi: 10.1007/s00432-015-1931-4

Wang, W. M. et al. 2017a. Synergistic role of Cul1 and c-Myc: Prognostic and predictive biomarkers in colorectal cancer. *Oncology Reports* 38(1), pp. 245-252. doi: 10.3892/or.2017.5671

Wang, W. M. et al. 2017b. Synergistic role between Cul1 and PARP1: prognostic and predictive biomarkers in colorectal cancer. *International Journal of Clinical and Experimental Medicine* 10(9), pp. 13992-+.

References

Wang, W. Y., Barratt, B. J., Clayton, D. G. and Todd, J. A. 2005. Genome-wide association studies: theoretical and practical concerns. *Nat Rev Genet* 6(2), pp. 109-118. doi: 10.1038/nrg1522

Wang, Z. et al. 2019. Association between Primary Tumor Location and Prognostic Survival in Synchronous Colorectal Liver Metastases after Surgical Treatment: A Retrospective Analysis of SEER Data. *J Cancer* 10(7), pp. 1593-1600. doi: 10.7150/jca.29294

Wasan, H. et al. 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), pp. 631-639. doi: 10.1016/s1470-2045(14)70106-8

Weisberg, J. F. a. S. 2019. *An {R} Companion to Applied Regression*. Third ed. Thousand Oaks CA: Sage.

Wenzel, J. et al. 2020. Loss of the nuclear Wnt pathway effector TCF7L2 promotes migration and invasion of human colorectal cancer cells. *Oncogene* 39(19), pp. 3893-3909. doi: 10.1038/s41388-020-1259-7

White, E. et al. 2004. VITamins And Lifestyle cohort study: Study design and characteristics of supplement users. *American Journal of Epidemiology* 159(1), pp. 83-93. doi: 10.1093/aje/kwh010

Wickham, H. et al. 2019. Welcome to the tidyverse. *Journal of Open Source Software* 4(43), p. 1686. doi: 10.21105/joss.01686

Widmer, C. et al. 2014. Further improvements to linear mixed models for genome-wide association studies. *Sci Rep* 4, p. 6874. doi: 10.1038/srep06874

References

Wigginton, J. E., Cutler, D. J. and Abecasis, G. R. 2005. A Note on Exact Tests of Hardy-Weinberg Equilibrium. *The American Journal of Human Genetics* 76(5), pp. 887-893. doi: https://doi.org/10.1086/429864

Willer, C. J. et al. 2010a. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* 26(18), pp. 2336-2337. doi: 10.1093/bioinformatics/btq419

Willer, C. J., Li, Y. and Abecasis, G. R. 2010b. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 26(17), pp. 2190-2191. doi: 10.1093/bioinformatics/btq340

Williams, C. S., Bernard, J. K., Demory Beckler, M., Almohazey, D., Washington, M. K., Smith, J. J. and Frey, M. R. 2015. ERBB4 is over-expressed in human colon cancer and enhances cellular transformation. *Carcinogenesis* 36(7), pp. 710-718. doi: 10.1093/carcin/bgv049

Wray, N. R., Lin, T., Austin, J., McGrath, J. J., Hickie, I. B., Murray, G. K. and Visscher, P. M. 2021. From Basic Science to Clinical Application of Polygenic Risk Scores: A Primer. *JAMA Psychiatry* 78(1), pp. 101-109. doi: 10.1001/jamapsychiatry.2020.3049

Xiao, R. and Boehnke, M. 2009. Quantifying and correcting for the winner's curse in genetic association studies. *Genet Epidemiol* 33(5), pp. 453-462. doi: 10.1002/gepi.20398

Yang, J. et al. 2016. Characteristics of Differently Located Colorectal Cancers Support Proximal and Distal Classification: A Population-Based Study of 57,847 Patients. *Plos One* 11(12), p. 12. doi: 10.1371/journal.pone.0167540

Yang, J. et al. 2011. Genomic inflation factors under polygenic inheritance. *Eur J Hum Genet* 19(7), pp. 807-812. doi: 10.1038/ejhg.2011.39

References

Yang, L. F. et al. 2020. Discrete functional and mechanistic roles of chromodomain Y-like 2 (CDYL2) transcript variants in breast cancer growth and metastasis. *Theranostics* 10(12), pp. 5242-5258. doi: 10.7150/thno.43744

Yarom, N. and Jonker, D. J. 2011. The role of the epidermal growth factor receptor in the mechanism and treatment of colorectal cancer. *Discov Med* 11(57), pp. 95-105.

Yurko, R., Roeder, K., Devlin, B. and G'Sell, M. 2021. H-MAGMA, inheriting a shaky statistical foundation, yields excess false positives. *Annals of Human Genetics* 85(3-4), pp. 97-100. doi: https://doi.org/10.1111/ahg.12412

Zhang, H. R. et al. 2022. Targeting PARP11 to avert immunosuppression and improve CAR T therapy in solid tumors. *Nature Cancer* 3(7), pp. 808-+. doi: 10.1038/s43018-022-00383-0

Zhang, L. N. et al. 2011. Functional SNP in the microRNA-367 binding site in the 3 ' UTR of the calcium channel ryanodine receptor gene 3 (RYR3) affects breast cancer risk and calcification. *Proceedings of the National Academy of Sciences of the United States of America* 108(33), pp. 13653-13658. doi: 10.1073/pnas.1103360108

Zhang, W. et al. 2019. IPO5 promotes the proliferation and tumourigenicity of colorectal cancer cells by mediating RASAL2 nuclear transportation. *Journal of Experimental & Clinical Cancer Research* 38(1), p. 296. doi: 10.1186/s13046-019-1290-0

Zhang, Y., Rajput, A., Jin, N. and Wang, J. 2020. Mechanisms of Immunosuppression in Colorectal Cancer. *Cancers* 12(12), p. 24. doi: 10.3390/cancers12123850

Zhao, S., Wu, W., Jiang, Z., Tang, F., Ding, L., Xu, W. and Ruan, L. 2022. Roles of ARID1A variations in colorectal cancer: a collaborative review. *Mol Med* 28(1), p. 42. doi: 10.1186/s10020-022-00469-6

References

Zheng, G. et al. 2019. Clinical validation of coexisting driver mutations in colorectal cancers. *Human Pathology* 86, pp. 12-20. doi: https://doi.org/10.1016/j.humpath.2018.11.014

Zhou, B. L., Zhu, W., Jiang, X. J. and Ren, C. P. 2019. RASAL2 Plays Inconsistent Roles in Different Cancers. *Frontiers in Oncology* 9, p. 6. doi: 10.3389/fonc.2019.01235

Zhu, H., He, G., Wang, Y., Hu, Y., Zhang, Z., Qian, X. and Wang, Y. 2019. Long intergenic noncoding RNA 00707 promotes colorectal cancer cell proliferation and metastasis by sponging miR-206. *Onco Targets Ther* 12, pp. 4331-4340. doi: 10.2147/ott.S198140

Zhu, Z. H. et al. 2016. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature Genetics* 48(5), pp. 481-+. doi: 10.1038/ng.3538

References

Appendices

Wills, C. et al. 2021. A genome-wide search for determinants of survival in 1926 patients with advanced colorectal cancer with follow-up in over 22,000 patients. European Journal of Cancer 159, pp. 247-258. doi: 10.1016/j.ejca.2021.09.047

Wills, C. et al. 2023. Germline variation in RASAL2 may predict survival in patients with RAS-activated colorectal cancer. Genes Chromosomes & Cancer 62(6), pp. 332-341. doi: 10.1002/gcc.23133